NEW TRENDS IN MULTIMEDIA AND NETWORK INFORMATION SYSTEMS

VISIT…

LANZAROTE Caliente.COM

Frontiers in Artificial Intelligence and Applications

FAIA covers all aspects of theoretical and applied artificial intelligence research in the form of monographs, doctoral dissertations, textbooks, handbooks and proceedings volumes. The FAIA series contains several sub-series, including "Information Modelling and Knowledge Bases" and "Knowledge-Based Intelligent Engineering Systems". It also includes the biennial ECAI, the European Conference on Artificial Intelligence, proceedings volumes, and other ECCAI – the European Coordinating Committee on Artificial Intelligence – sponsored publications. An editorial panel of internationally well-known scholars is appointed to provide a high quality selection.

Volume 181

Recently published in this series

Vol. 180. M. Virvou and T. Nakamura (Eds.), Knowledge-Based Software Engineering – Proceedings of the Eighth Joint Conference on Knowledge-Based Software Engineering
Vol. 179. A. Cesta and N. Fakotakis (Eds.), STAIRS 2008 – Proceedings of the Fourth Starting AI Researchers' Symposium
Vol. 178. M. Ghallab et al. (Eds.), ECAI 2008 – 18th European Conference on Artificial Intelligence
Vol. 177. C. Soares et al. (Eds.), Applications of Data Mining in E-Business and Finance
Vol. 176. P. Zaraté et al. (Eds.), Collaborative Decision Making: Perspectives and Challenges
Vol. 175. A. Briggle, K. Waelbers and P.A.E. Brey (Eds.), Current Issues in Computing and Philosophy
Vol. 174. S. Borgo and L. Lesmo (Eds.), Formal Ontologies Meet Industry
Vol. 173. A. Holst et al. (Eds.), Tenth Scandinavian Conference on Artificial Intelligence – SCAI 2008
Vol. 172. Ph. Besnard et al. (Eds.), Computational Models of Argument – Proceedings of COMMA 2008
Vol. 171. P. Wang et al. (Eds.), Artificial General Intelligence 2008 – Proceedings of the First AGI Conference
Vol. 170. J.D. Velásquez and V. Palade, Adaptive Web Sites – A Knowledge Extraction from Web Data Approach
Vol. 169. C. Branki et al. (Eds.), Techniques and Applications for Mobile Commerce – Proceedings of TAMoCo 2008
Vol. 168. C. Riggelsen, Approximation Methods for Efficient Learning of Bayesian Networks

ISSN 0922-6389

New Trends in Multimedia and Network Information Systems

Edited by

Aleksander Zgrzywa

Wrocław University of Technology, Wrocław, Poland

Kazimierz Choroś

Wrocław University of Technology, Wrocław, Poland

and

Andrzej Siemiński

Wrocław University of Technology, Wrocław, Poland

IOS Press

Amsterdam • Berlin • Oxford • Tokyo • Washington, DC

ISBN 978-1-58603-904-2
Library of Congress Control Number: 2008933359

Publisher
IOS Press
Nieuwe Hemweg 6B
1013 BG Amsterdam
Netherlands
fax: +31 20 687 0019
e-mail: order@iospress.nl

Distributor in the UK and Ireland
Gazelle Books Services Ltd.
White Cross Mills
Hightown
Lancaster LA1 4XS
United Kingdom
fax: +44 1524 63232
e-mail: sales@gazellebooks.co.uk

Distributor in the USA and Canada
IOS Press, Inc.
4502 Rachael Manor Drive
Fairfax, VA 22032
USA
fax: +1 703 323 3668
e-mail: iosbooks@iospress.com

PRINTED IN THE NETHERLANDS

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

New Trends in Multimedia and Network Information Systems

Preface

We want to present to our readers this new monograph on the current trends in Multimedia and Network Information Systems. It discusses a very broad scope of subject matters including multimedia systems in their widest sense, Web systems, and network technologies. The monograph also includes texts devoted to more traditional information systems that draw on the experience of the Multimedia and Network Systems. Each of the discussed research trends is considered from both theoretical and practical viewpoints. Imposing a clear-cut classification for such a diverse research area is not an easy task.

The challenge is even greater due to the fact that in this book we tried to focus on the most topical research work of scientists from all over the world. The studies are original and have not been published anywhere else. In our opinion the chapters represent the dominant advances in computer information systems. It is worth emphasizing, that in most cases the research work relies heavily on the achievements and techniques developed originally in the area of Artificial Intelligence.

As a result, we have divided the monograph content into four major parts:

1. Multimedia Information Technology.
2. Data Processing in Information Systems.
3. Information System Applications.
4. Web Systems and Network Technologies.

Each of the parts covers a couple of chapters on detailed subject fields that comprise the area of its title.

We do hope that we have managed to collect and systematize the scientific knowledge on such a diverse field. We will be very pleased if this book inspires the research community working on Multimedia and Network Information Systems. If so, it will have achieved the goal that motivated the authors, reviewers, and editors.

Aleksander Zgrzywa
Kazimierz Choroś
Andrzej Siemiński

This page intentionally left blank

Reviewers

Witold ABRAMOWICZ	Poznań University of Economics, Poland
Costin BADICA	University of Craiova, Romania
Leszek BORZEMSKI	Wrocław University of Technology, Poland
Noelle CARBONELL	Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA), Nancy, France
Andrzej CZYŻEWSKI	Gdańsk University of Technology, Poland
Paul DAVIDSSON	Blekinge Institute of Technology, Sweden
Luminita DUMITRIU	Dunarea de Jos University, Romania
Bogdan GABRYŚ	Bournemouth University, UK
Przemysław KAZIENKO	Wrocław University of Technology, Poland
Paul KLIMSA	Technische Universitat Ilmenau, Germany
Elżbieta KUKLA	Wrocław University of Technology, Poland
Zygmunt MAZUR	Wrocław University of Technology, Poland
Tadeusz MORZY	Poznań University of Technology, Poland
Anton NIJHOLT	University of Twente, Enschede, The Netherlands
Toyoaki NISHIDA	Kyoto University, Japan
Tarkko OKSALA	Helsinki University of Technology, Finland
Maciej PIASECKI	Wrocław University of Technology, Poland
Janusz SOBECKI	Wrocław University of Technology, Poland
Stanisław SZPAKOWICZ	University of Ottawa, Canada and Institute of Computer Science, Polish Academy of Sciences, Poland
Bogdan TRAWIŃSKI	Wrocław University of Technology, Poland
Wojciech ZAMOJSKI	Wrocław University of Technology, Poland

This page intentionally left blank

Contents

Information System Applications

Web Systems and Network Technologies

Multimedia Information Technology

This page intentionally left blank

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-3

Estimation of Similarity between Colour Images

Paweł BENECKI and Adam ŚWITOŃSKI
Institute of Informatics, Silesian University of Technology in Gliwice, Poland
Akademicka 16, 44-100 Gliwice, Poland

Abstract. The paper describes and compares content-based global and local feature image retrieval techniques, which can be applied to estimate similarity between colour images. The database of 150 colour images with manually specified model similarity of each pair was prepared. The comparison of the analyzed methods was based on examining model and calculated similarities, which should be alike.

Keywords. multimedia databases, colour image processing, image retrieval, similarity of images, histogram analysis, image global features, image local features

Introduction

Content based image retrieval has already shown its usefulness in many applications. Many systems which retrieve images containing given person faces, fingerprints or objects have been implemented. A lot of them do not use colour information contained in images, although colour is very important in human vision. Furthermore, in many cases these systems are dedicated only to a single task – for example face recognition. The problem to find similar images to a particular one often occurs in multimedia databases of colour images. Unfortunately, there is not one solution for all types of images. It happens because similarity can have different meaning and it depends on interpretation of images. In some cases global colour statistics are enough for an estimation of similarity but in others, local features have to be analyzed. Searching in image databases is a challenge, because the results must be reliable and it should be computed fast.

In the paper we focus on the methods of estimation of similarity between colour images using colour histograms in RGB and HSV colour spaces, central moments of histograms in these spaces and local SIFT descriptor [7]. Additionally, for central moments we consider image division into regions to provide pseudo-localization of features and we compare the results with those computed for the whole image. We analyze different ways of computing distances between histograms. We have implemented all described algorithms and tested them for the prepared test image database with manually specified model similarity. We propose two ways of comparison methods, in which we expect to receive the same ordering of images by model and calculated similarity to a given image.

1. Global Colour Features

1.1. Histograms

1.1.1. RGB Colour space

Full RGB image histogram should be 3-dimensional, which would make any operations extremely complex - memory and computational expensive. That is why to produce useful feature vector colour quantization has to be applied. It turns multidimensional histogram into one-dimensional feature vector. As in QBIC [2] system, each value of colour channel is reduced to 4 possible values. This results in 64 base colours. Before the quantization, the picture is resized not to be greater than 200 pixels in longer dimension for reducing computational complexity. Furthermore, feature vector is normalized, which makes it independent of image resolution.

1.1.2. HSV Colour space

HSV (Hue, Saturation, Value) colour space offers more perceptually uniform representation of colours [1], [3] than machine-oriented RGB. Quantization of colours is done in way which was proposed in VisualSEEK system [1], [3]. Authors claim that their algorithm reduces number of colours retaining colour-robustness of image. Prior to quantization, median filter is applied on each colour channel to eliminate outliers. Hue channel is partitioned into 18 values, saturation and value into 3 values, which gives 162 colours. For undefined hue there are 4 more greys. This gives 166 colours in total.

1.2. Central moments of histograms

A new kind of feature vector characterizing image containing first three central moments of histograms of each colour channels: mean, variance and skewness is proposed in [4]. According to authors, this should be sufficient to produce good quality feature vectors. The central moments are defined by Eq. 1, Eq. 2 and Eq. 3.

$$E_i = \frac{1}{N}\sum_P C_i(P) \tag{1}$$

$$\sigma_i = \sqrt{\frac{1}{N}\sum_P \left(C_i(P) - E_I\right)^2} \tag{2}$$

$$s_i = \sqrt[3]{\frac{1}{N}\sum_P \left(C_i(P) - E_I\right)^3} \tag{3}$$

$C_i(p)$ represents value of i colour channel in point P.

Such 9-element feature vectors can be computed for whole images or for its every selected region, which provides localization of features giving better search results [4]. Figure 1 shows image division into regions proposed in [4].

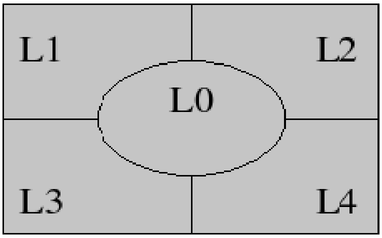


Figure 1. Regions in Stricker-Dimai method

What is more, in [4] fuzziness is introduced to computation – the membership function which has value of 1 for the pixel inside region and value of R for point outside region:

$$R = \frac{1}{2}\left(\cos\left(\frac{d\pi}{r}\right) + 1\right) \tag{4}$$

where r represents fuzzy radius and d is distance to region. This results in five 9-element feature vectors for a single image. Details can be found in [4].

1.3. Comparison of global features

For feature vectors described in previous sections we define functions which provide measure of distance between them.

1.3.1. Distance between histograms

Basic measure is the Euclidean distance between feature vectors H_i and H_2:

$$d_{H_1,H_2} = \sqrt{\sum_{i=1}^{N} \left(h_1^{(i)} - h_2^{(i)}\right)^2} \tag{5}$$

Its drawback is that it does not consider small shifts between colours, like light orange and orange, in two histograms.

Another way of calculating distance between histograms is cross-distance function [5], [2]:

$$d_{H_1,H_2} = (H_1 - H_2)^T A(H_1 - H_2) \tag{6}$$

where A is colour similarity matrix in specified colour space. Elements of matrix are defined in [5] as:

$$a_{ij} = 1 - \frac{d_{ij}}{\max_{ij}(d_{ij})} \tag{7}$$

where d_{ij} is distance between colours and max(d) is the maximum distance between colours in the specified colour space. For RGB colour space distance is defined as:

$$d_{C1,C2}^{RGB} = \sqrt{(R_{C1} - R_{C2})^2 + (G_{C1} - G_{C2})^2 + (B_{C1} - B_{C2})^2} \tag{8}$$

and for HSV this is ([1]):

$$d_{C1,C2}^{HSV} = \sqrt{\begin{matrix}(V_{C1} - V_{C2})^2 + (S_{C1}\cos H_{C1} - S_{C2}\cos H_{C2})^2 + \\ (S_{C1}\sin H_{C1} - S_{C2}\sin H_{C2})^2\end{matrix}} \tag{9}$$

One more method of computing distance – histogram intersection is proposed in [6]:

$$d(H_1, H_2) = 1 - H_1 \cap H_2 = 1 - \sum_{i=1}^{N} \min(h_1^{(i)} - h_2^{(i)}) \tag{10}$$

1.3.2. Distance between histogram central moments feature vectors

The distances between feature vectors consisting of central moments of histograms are defined in [4]. For feature vectors calculated on whole image the function is following:

$$d(H,J) = \sum_{i=1}^{c} \left(w_{i1} \left| E_i^H - E_i^J \right| + w_{i2} \left| \sigma_i^H - \sigma_i^J \right| + w_{i3} \left| s_i^H - s_i^J \right| \right) \tag{11}$$

where values of w are arbitrarily defined weights. Authors of [4] propose example values, which emphasis the matching the first and the second moments for the hue and the saturation:

$$\begin{array}{cccc} & E & \sigma & s \\ H & 3 & 3 & 1 \\ S & 2 & 1 & 1 \\ V & 1 & 1 & 1 \end{array}$$

If we divide image into regions, we can treat central area in a special way. Interesting objects are often located in the centre of an image. Thus we can pay more attention to it.

$$d(H,J) = S_0 R_0^{tot} d_{0,0}(H,J) + \min_{f \in T_{90}} \sum_{i_1=i_2=1}^{4} S_{i_1 i} R_{i_1}^{tot} d_{i_1, f(i_2)}(H,J) \qquad (12)$$

The iteration f means rotation over centre of image by 0º, 90º, 180º, 270º and provides rotation invariance. The angle in which the distance is summarily smallest is chosen. Weights S determine how important each region should be in summary distance. For example, when we want to consider only the centre of image we set S_0 to no-zero value and the others S_i to zero.

2. Local Features – SIFT

In our investigation we use SIFT (Scale Invariant Image Transform) presented in [7] to detect keypoints on image and describe regions around them. The SIFT algorithm detects keypoints regardless of scale in which they are located. Detailed description of algorithm can be found in [7], here we present only the main concepts.

To detect keypoints a "pyramid" of image sequences is built - as in Figure 2. On each level of the pyramid there are images which have the same dimensions. Authors call such sequence an "octave". In every octave images are subsequently Gaussian blurred:

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y) \qquad (13)$$

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{\sigma^2}} \qquad (14)$$

Value of σ is multiplied by $k=\sqrt{2}$ every time the image is blurred. The differences of Gaussians (DoG) are computed as:

$$D(x,y,\sigma)=L(x,y,\sigma)-L(x,y,k\sigma) \tag{15}$$

The number of images in octave is chosen according to [7] to be 5, so we obtain four DoG matrices, as shown on Figure 2.

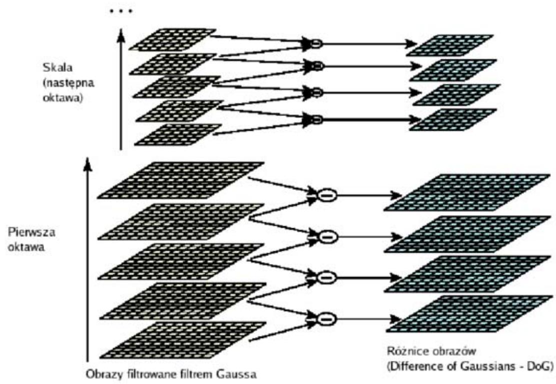


Figure 2. Building sequence of images [7]

We search for extrema in every DoG having both neighbouring layers. The extreme point must have either the greatest or the smallest value from: 8 points on the same layer around it, 9 points on the above layer and 9 points on the underlying layer. This is illustrated on Figure 3.

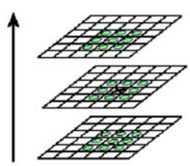


Figure 3. Extrema localization [7]

Some subsequent operations are done to eliminate part of the extrema.

In the next step the main orientation and magnitude m of image gradient around the extreme point are found. In the result we can obtain rotation invariance of the keypoint.

Then, SIFT descriptors for every keypoint are computed. The descriptor is a set of gradient histograms calculated for the every of sixteen keypoint surroundings, what is shown on Figure 4.

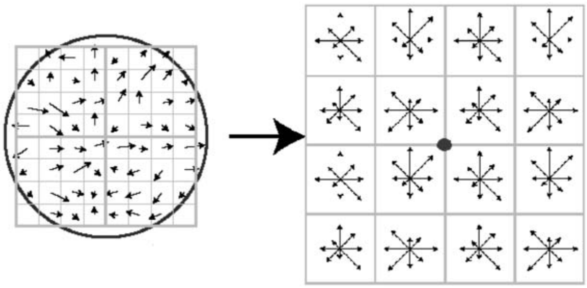


Figure 4. SIFT Descriptor building [7]

Finally, feature vector consists of 128 values: 16 areas and 8 bins in every histogram.

To match points from two images we define distance between descriptors, which is basic Euclidean measure:

$$d(D_1, D_2) = \sqrt{\sum_{i=1}^{128} \left(d_1^{(i)} - d_2^{(i)}\right)^2} \qquad (16)$$

For every keypoint from the test image we find two nearest keypoints from the second image. If the quotient of these distances is greater than 0.8 – the match is eliminated.

In the following step Hough transform [8] is applied to the set of matches to group them by the main orientation, scale and translation. Received groups are described by parameters of affine transform, which transforms the keypoints of every match within a group on the first image to the keypoints on the second one. Finally imprecisely transformed matches are rejected by iterative algorithms. If at the end more than three matches stay in a group, the group is accepted and objects of analyzed images are considered to be matched.

In [9] authors define mean error, which can be used as distance metric between set of matches of the group and estimate the detection accuracy:

$$e = \sqrt{\frac{2\|Ax - b\|^2}{r - 4}} \qquad (17)$$

Detail explanation of matrices ***A***, ***b***, ***x*** and parameter ***r*** can be found in [7] and [9].

The above algorithms do not use information about colours. Although, there is a colour version of SIFT proposed in [10], in our investigations we have used “hybrid” solution – keypoints are found and described independently for every colour channel, so colour information is utilized.

3. Test Database

We have built example test database of real photo images. We have assumed to compare algorithms with human similarity-feeling in general image database. That is why we have not bounded to any kind of images. As in most real databases we have chosen pictures containing the same or similar objects, with similar colours and pictures which are different from the rest. The prepared database consists of 150 images and the main groups of pictures are as follows:

- photos of the same building, all taken from similar perspective and scale, but in different light and weather conditions,
- photos of vehicles of the same colour, similar mountain landscapes and flowers,
- photos of a cat in different backgrounds,
- photos of objects in similar scale but different configurations - some of them contain only objects,
- other photos, not similar to the rest by colour nor by contained objects.

Example images from the database are shown on Figure 5.

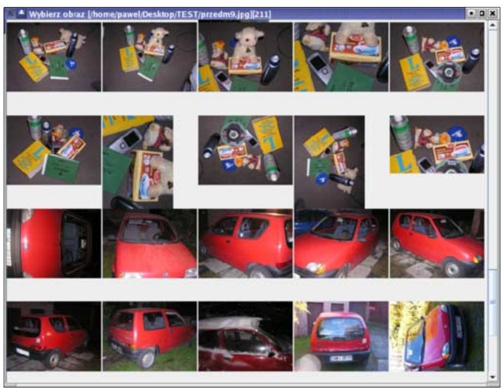


Figure 5. Fragment of the test image database

We have manually assigned to each pair of pictures in the database a degree of similarity – in scale from 0-5. 0 means that pictures are not similar at all, and 5 – that images are almost the same. Similarity degrees are used in the comparison technique described below.

4. Tests and results

We have built test application which can be used as image database system with a functionality of searching similar images. The data is stored on the level of relational database server and image retrieval algorithms are implemented on the client side.

The basic function of our system is creation of a query which returns an image set ordered by similarity to the particular image. The method of estimating similarities can be specified.

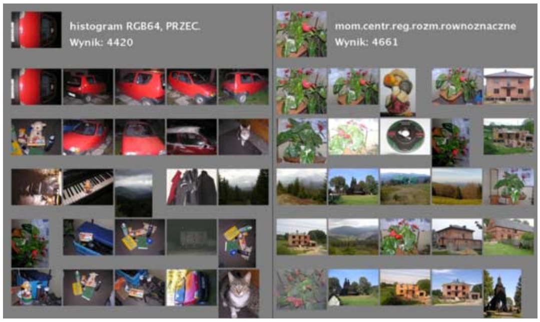


Figure 6. Example result list

In Figure 6 there are snapshots of our application with query results. Searched image is at the top and below there is a list of images beginning from the most similar one.

5. Evaluation of the algorithms

To evaluate accurately the results we propose algorithm which generates a number called "result quality factor" (RQF) which characterizes the result. Distances calculated by the algorithms produce sort order of the result list. It should be similar to the order specified by the model similarity. We assume that the position on the result list of images and similarity degree assigned beforehand, are the main criteria. We expect that the most similar pictures are at the beginning of the result list. The result quality factor can be defined in two ways:

$$RQF = \sum_{i=1}^{N} (n-i) sim(m, t_i) \tag{18}$$

$$RQF = \sum_{i=1}^{N} sim(m, t_i) \frac{1}{\sqrt{2\pi}} e^{\frac{i^2}{2*400}} \tag{19}$$

where $\boldsymbol{m}$ is the model image, $\boldsymbol{T}$ is ordered result list of images, $\boldsymbol{i}$ is position on result list, $\boldsymbol{t_i}$ is image from the list, $\boldsymbol{n}$ is size of result list and ***sim(a,b)*** is model similarity of images defined manually.

The first equation defines quality factor in a linear way – uses only position on the result list. The second one promotes images at the beginning of the list. It imitates a natural process of a human being, who is usually more interested in the most similar images, ignoring the less alike ones. The comparison of linear and Gausian RQF weights is illustrated in Figure 7.

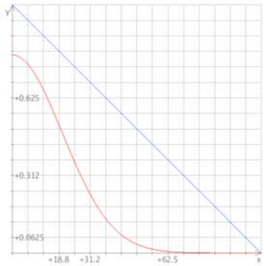


Figure 7. Comparison of linear and Gaussian RQF weights

The closer sort orders specified by computed and model similarities, the greater values of both **RQFs.**

We have chosen randomly and evenly for each database group 15 images and executed for them all the described methods:

- three methods of comparing histograms: Euclidean distance, cross-distance and intersection, for HSV and RGB colour spaces,
- central-moments feature vector calculated for whole image,
- central-moments feature vector calculated for regions – in fuzzy and not fuzzy version,
- central-moments feature vector calculated for regions where central region is 5 times more important than the others -in fuzzy and not fuzzy version,
- SIFT algorithm.

What is more the results were normalized by the factor equal to RQF for the best hypothetical search result – the ideal case of sorting. The final presented RQFs are average results of a given methods calculated for all tested images.

Numerical results are shown in Table 1 and its graphical presentation in Figure 8 and Figure 9.

Table 1. Results of algorithms for linear and Gaussian comparison method

	Algorithm	RQF (linear)	RQF (Gaussian)
1	RGB histogram, Euclidean distance	0.42	0.34
2	RGB histogram, intersection	0.44	0.37
3	RGB histogram, cross-distance	0.43	0.36
4	HSV histogram, Euclidean distance	0.41	0.35
5	HSV histogram, intersection	0.45	0.38
6	HSV histogram, cross-distance	0.42	0.35
7	Central moments – whole image	0.42	0.34
8	Central moments – fuzzy regions	0.41	0.33
9	Central moments – fuzzy regions, central area 5 times more important	0.42	0.34
10	Central moments – not fuzzy regions	0.41	0.34
11	Central moments – not fuzzy regions, central area 5 times more important	0.42	0.34
12	SIFT	0.33	0.20

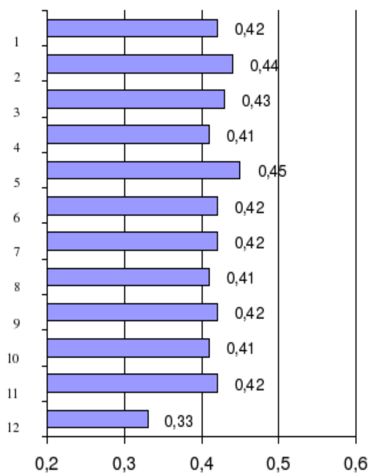


Figure 8. Results for linear comparison method

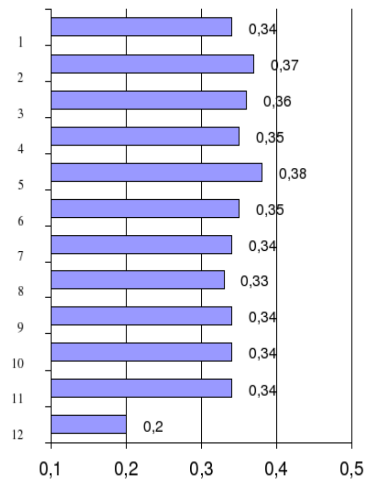


Figure 9. Results for Gaussian comparison method

The comparison of HSV and RGB colour spaces is presented in Table 2. It contains values for a histogram retrieval results – average of quality factors for each distance.

Table 2. Results for RGB and HSV

	RQF (linear)	RQF (Gaussian)
RGB	0.43	0.36
HSV	0.43	0.36

6. Conclusions

On analyzing the results presented in the previous section – Table 1, Table 2, Figure 8 and Figure 9, we have come to the following conclusions.

The most effective methods of estimation similarity between images for the test database are based on the histogram comparison. Surprisingly, the histogram intersection - method that is the most simple and the least computationally expensive is also the most effective. More sophisticated methods, like cross-distance, although better than basic Euclidean method, are less accurate yet. Estimation of similarity by the central moments is only a slightly worse than by histogram analysis. In this case the simplest method – the central moments calculated for whole image, is also the most accurate.

For both RQFs best results are achieved by HSV colour space with intersection distance measure. In spite of that, average results for HSV and RGB colour spaces are almost the same.

In general, all methods using global colour features give satisfactory comparable results and are useful in searching images in large databases.

Another subject is poor result of SIFT algorithm. This method is excellent for localization of objects when they are present on both images. If they are not, it can sometimes find wrongly matched keypoints. Examples of both cases are shown on Figure 10. What is more, our database consists of photos with the same objects viewed from different sides, for example the front and the back of a car. In such a situation there is no global affine transform for matched keypoints, because they do not correspond with each other in both images and localization fails. The last reason for poor SIFT result is the mean error function for the best group, which estimates similarity. In fact, we distinguish two cases – objects are localized, which means images are similar or not. The mean error cannot be calculated if all matched keypoints are rejected, but that does not always mean that images are not similar at all.

Figure 10. Results of SIFT algorithm – matched keypoints are connected

Another issue is computational complexity of the algorithms, which results in the execution time of a search query. For the system with dominant search operations, we propose storing of the global features and the SIFT keypoints in the database rather than calculating them separately for every search. This reduces greatly time of computation. Each of the global-feature oriented algorithms needs in such case only a couple of milliseconds to process an image. Thus, for databases containing thousands of pictures results are completed in a few seconds, using only one processing unit. The SIFT is slow indeed. Extracting keypoints from an image consumes a several seconds and matching them takes another few more. That is why SIFT is not a satisfactory solution for content base image retrieval in large image databases.

Our system has been implemented in JAVA language and tests have been made on the PC with Pentium IV 3.0 GHz processor and 512 MB of RAM.

References

[1] J. R. Smith and S.-F. Chang, VisualSEEk – a fully automated content – based image query system, *Proc. ACM Conf. Multimedia*, ACM Press, New York, 1996.

[2] E. Faloutsos, N. Flickner, B. Petkovitz, Efficent and Effective Quering by Image Content, *Journal of Intelligent Information Systems*, 1994.

[3] J. R. Smith and S.-F. Chang, Single Color Extraction and Image Query, *Proceedings of International Conference on Image Processing*, 1995.

[4] M. Stricker and A. Dimai, Color indexing with weak spatial constraints, *Storage and Retrieval for Image and Video Databases ({SPIE})*, 1996.

[5] R. Schettini, G. Ciocca, and S. Zuffi, *A survey of methds for colour image indexing and retrieval in image databases*, Instituto Technologie Informatiche Multimediali, Milano.

[6] S. M. Lee, J. H. Xin, and S. Westland, Evaluation of Image Similarity by Histogram Intersection, *Color Research and Application*, John Wiley & Sons ltd., 2005.

[7] D. Lowe, Object Recognition from Local Scale-Invariant Features, *International Conference on Computer Vision*, Corfu, Greece (September 1999).

[8] L. Chmielewski, *Nakładanie obrazów metodą transformaty Hougha*, Instytut Podstawowych Problemów Techniki PAN, 2003.

[9] D. Lowe, Local feature view clustering for 3D object recognition, *CVPR Proc.*, Springer, 2001, 682-688.

[10] A.E. Abdel-Hakim and A.A. Farag, CSIFT: A SIFT Descriptor with Color Invariant Characteristics, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-16

Image Identification Based on the Pupil Size Analysis During Human-Computer Interaction

Janusz SOBECKI
Institute of Applied Informatics, Wrocław University of Technology
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
e-mail: sobecki@pwr.wroc.pl

Abstract. A method for image identification based on the pupil size analysis during human-computer interaction will be presented. The data was gathered during the experiment on 3D interactive representation of fairly large picture collections which facilitates browsing through unstructured sets of icons or pictures. The data was gathered using ASL gaze-tracking device. A group of 18 users took part in the experiment. The obtained gaze tracking results have shown that recorded pupil size patterns may indicate the fact of the presented image identification.

Keywords: gaze-tracking, image identification, pupil size

Introduction

Nowadays, there are quite a lot of research done in the area of new interaction styles, which are far away from standard direct graphical manipulation [7] and reach even brain-computer interface [8]. It is obvious that such new methods require an application of sophisticated devices and also some physical interference with the human body, which could be at least inconvenient for the users.

There are, however, other possible technologies that could be applied in modern human-computer interfaces. One of which is gaze-tracking that could be made using for example ASL devices such as head mounted optics (HMO) and Pan/Tilt (P/T) [1,2] and Head Tracker (HT). These devices together with appropriate software such as EyeTrack 5000 or EyeTrack 6000 and GTAnaly are able to track several parameters concerning user interaction with computer application, such as: gaze fixation points coordinates, pupil size and head position together with time stamps, cursor movements and keyboard actions.

In this paper we present the method for identification of the recognition of the specified image by the user. The identification is made using special gaze-tracking devices and especially the pupil size and head position. The data was gathered during the experiments with usability evaluation study of 3D versus 2D presentations of large photo collections.

This method may be applied in many different interaction systems for which usage of other input devices is impossible, very difficult or increases the interface the user-computer interaction. One of the most obvious applications is user-computer

interaction of disabled, then, we should encounter application in all the situations where users have hands busy with other duties, such as car driving or operating some machines.

The paper is organized as follows: in the second paragraph the ASL gaze-tracking technology is presented shortly, in the third paragraph the experiment on usability evaluation study of 3D versus 2D presentations of large photo collections is presented, in the fourth the method of identification of the image recognition is presented and finally in the fifth paragraph the conclusion remarks are given.

1. Gaze Tracking Analysis and Pupil Size

In the experiment we used ASL Model 501 head mounted eye tracker (HMO) [2] that is designed to accurately measure a person's eye line of gaze with respect to his/her head movements measured by Head Tracker (HT) and ASL's EYEHEAD integration software (EHS option), the model 501 eye tracker can also measure a person's line of gaze with respect to stationary surfaces in the environment. The eye measurement is displayed as a cursor superimposed on the image from a head mounted scene camera. The measurements may be recorded digitally on the eye tracker Interface PC, or exported as a real time serial data stream to an external device. As we used the HT and ASL's EYEHEAD integration software (EHS option), the model 501 eye tracker also measures a person's line of gaze with respect to stationary surfaces in the environment, The PC serves as both the user interface with the eye tracker and as a digital data recording device.

The system works as follows: the user eye is illuminated by the a near infrared beam source, the illumination beam and the eye image are reflected by the semi-reflected mirror attached to the helmet, the eye image is recorded by eye camera, the scene camera may be focused on the scene being viewed by the subject, i.e. computer monitor. All the elements are helmet (HMO) mounted.

A minimum system configuration for any computer application analysis requires the following devices: Model 5000 Control Unit, Helmet or Headband mounted optics including eye camera optics module (HMO) and a head mounted scene camera, Two video monitors (one for the eye image and one for the scene image), control PC, application PC, necessary cables connecting these devices through specified interfaces and software for data gathering Eye-track 5000, and optionally for its analysis GTAnaly [1], which is a multipurpose tool allowing collection, analysis, and visualization of captured eye-gaze and computer data.

The eye tracking starts with the pupil and CR (cornea reflection) recognition that is performed using edge detection logic and shown on the video eye monitor as white and black circles with corresponding cross hairs indicating the center respectively. To achieve proper pupil and CR discrimination one should calibrate the pupil discriminator level until both centers are properly recognized as the subject looks about the field of view of interest.

After collecting the data with GTAnaly software we can analyze and visualize the obtained results in the following forms: gaze trail, contour, look zones 3D, look zones order and pupil graph. We can also combine several views in a single image. We can use however also other proprietary software for gaze tracking analysis.

The gaze tracking analysis is based on the assumption that the center of the fixation of the eye indicates the center of the attention in the given time [4]. As the

pupil size is our main interest here, it is worth noting what it indicates. The main purpose of the pupil that is located in the centre of the iris of the eye is controlling the amount of light entering the eye. The pupil size reflects also state of the human body and mind, it may change with changes of the user mental activity as well as when the user is fearful, angry, in pain, in love, or under the influence of drugs. For example it was observed that there is a close correlation between mental activity and the difficulty of the problem, which results in increase in pupil size when the correct answered are given and in the corresponding decrease for in pupil size for wrong replies.

2. Experiment Description

The data gathered for analysis was collected during experiments [3] with large photo collection retrieval described. In this experiment photos in form of digital images were retrieved by applying visualization techniques designed for data mountains (e.g., treemaps) to the display of photo thumbnails [4],[5]. The main goal of the experiments was comparison of the two basic large photo collection presentation approaches [6]: 3D (three dimension) and 2D (two dimension) visualizations.

However, only data from standard 2D view image identification were analyzed (see fig. 1). In this part of the experiment 18 users of different age (from 21 to 30) and gender were asked to locate the known image, zoom it and confirm the localization of the proper photo. Almost in all cases users identified the proper photo, however, we encountered at least one error.

Collected data used in our study contained traces of participants’ interactions with the 2D view that were recorded together with their eye movements and pupil size by using head-mounted ASL 501 eye tracker, during the search for a visually familiar photo the location of which in the collection representation was also familiar to participants. These recordings were made in specified constant time intervals that were equal 0.16 s.

Fig. 1. 2D presentation of large collection of photos [3]

3. Image Identification Method Description

In the experiment presented in section 3 the pupil size data was recorded but was not taken into consideration. It was noticed however that some pupil size patterns may be observed when users open correct pictures and some other when they open wrong ones.

In this experiment gaze-tracking data on 18 different users was gathered [3], however only half of them were taken for the thorough analysis. We reduced the number of analyzed users to balance the number of users who have at least one wrong picture and all the correct ones. The users have different age and gender, so in consequence they exhibit different reactions to the stimuli. First, they have different pupil size, so we analyzed the relational changes that were calculated in the following way:

$$rps = \frac{ps - mps}{mps},$$

where: *rps* denotes relational pupil size of the user, *ps* actual pupil size of this user, and *mps* minimal pupil size measured for this user.

The gaze tracker records also sizes as small as zero that are recorded during closing the eye by the eyelid and are not valid minimal pupil sizes. In that case the *rps* is set to 0. Second, the distance from the monitor was also different, so we analyzed the relational changes that were calculated in the following way:

$$rd = \frac{d - md}{md},$$

where: *rd* denotes relational distance of the head of the user to the monitor, *d* actual distance, and *md* minimal distance respectively.

Having these two measures we analyzed their values when users opened the image for identification in the 2D image interaction. The users may open the correct or the incorrect image. They confirm it by clicking the specified buttons. We checked the *rps* and *rd* values in a period of time when the user viewed the opened image and we observed specified patters. We observed that when the opened image is the correct one then the *rps* changes rapidly for a short time and then returns to the previous values. These changes mean that the pupil becomes bigger for a while, but for some users the changes are opposite or users close their eyes for a moment. As for the second value, the related head distance *rd* in many cases also changes, but these changes, the head gets closer to the monitor, last longer. The figure 2 shows the related pupil size (Series 1) and related head distance (Series 2) changes in analyzed 73 time intervals for 0.166485 seconds each for correct image.

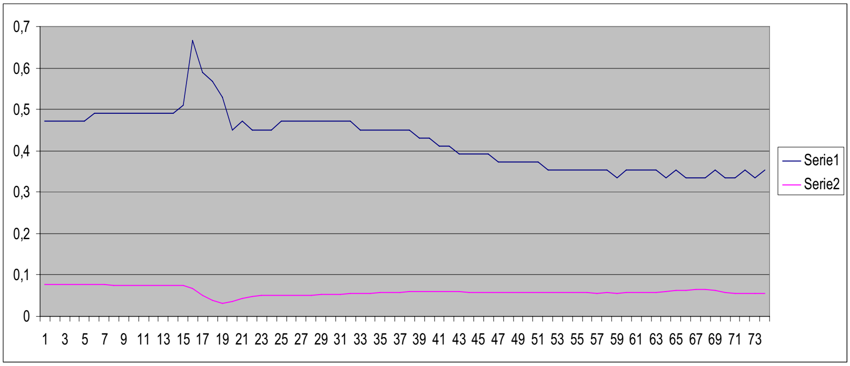


Fig. 2. Relative pupil size and distance for subject 5 image img4_352 - target 1

However, several users blinked the eye (sometimes even more than once) during the identification, so the relative pupil size had different diagram (see fig. 3) than in the previous case. We consider that blinking (see relative pupil size values about value 289 of the time interval) may also be good indicator for correct image identification.

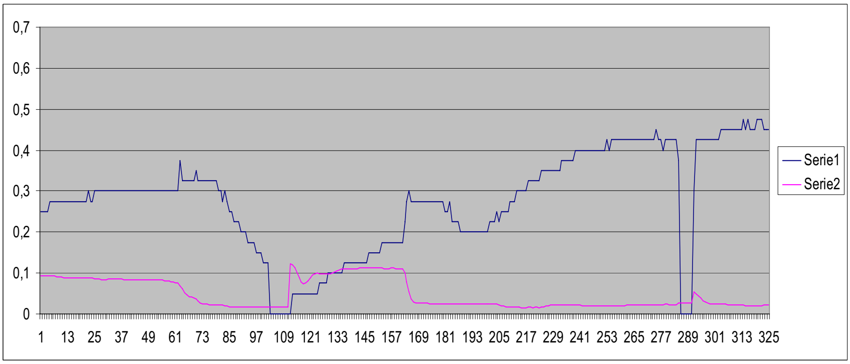


Fig. 3. Relative pupil size and distance for subject 11 image img4_352 - target 1

In case the image that was opened by the user and was identified as a wrong one, we noticed that there were not any rapid changes in relative pupil size and distance as presented in fig. 4. We can observe some changes in the relative pupil size but not so rapid. Also the direction of change is opposite, it is getting smaller. We also do not observe any significant change in the head distance.

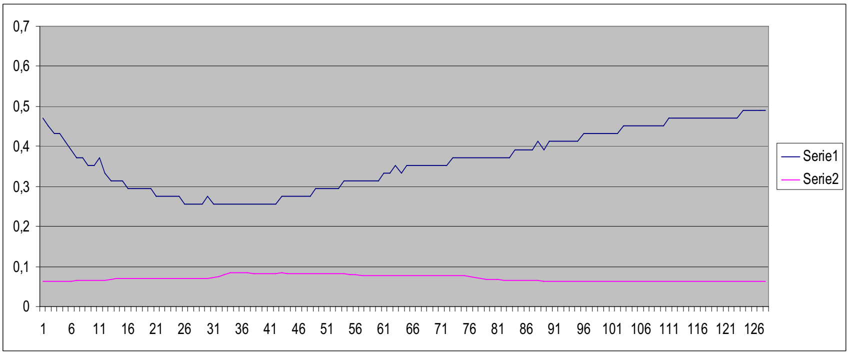


Fig. 4. Relative pupil size and distance for subject 5 image img3_715 - not target 3

For some users we noticed that for wrong images we encountered series of rapid pupil size changes in the whole analyzed period of identification (see fig. 4). However, in this case we did not observed significant changes in the relative distance from the monitor.

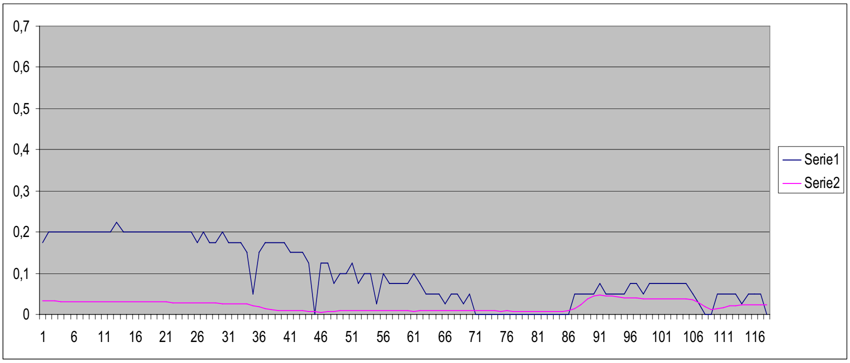


Fig. 4. Relative pupil size and distance for user 13 image img3_868 - not target 4

The final results of our study were as follows. The 9 users who were analyzed viewed 63 images, out of which 53 was correct ones and 10 incorrect. Using the above mention method of identification we were able to positively identify ~90% of correct pictures with ~7% of false positives and ~55% of positively identified incorrect pictures and ~44% of false positives for these pictures. The rates for correct pictures are quite satisfactory, however, the rates for incorrect ones are rather poor.

4. Summary

We presented a pretty simple method for identification of correct and incorrect images based on the gaze-tracking data containing information on the user pupil size and head distance to the screen. We observed that opening of a correct picture is accompanied with rather rapid change in the pupil size, most often rapid increase and then decrease

of the relative pupil size (a hill shape), rapid separate decrease or increase of this value, or eye blink. Quite often (in about 73%) the opening of the positive image is accompanied with decrease or increase of the relative head distance to the monitor. The identification of the incorrect pictures is less effective, however, most of the errors were made by only one of the user (U5) and in one case the other user (U1). The false positive identification of the incorrect picture may be explained by the fact that the user was sure it is correct one.

The obtained results are promising, however, to obtain more reliable results we must make more and more specialized experiments. We presented only some conclusions based on the experiment designed for different hypothesis, however, making such experimentation is quite time consuming and pretty expensive because of the utilization of the specialized equipment. We observed also different patterns for each user, so it is possible that identification of correct and incorrect pictures should be tuned for each user separately.

References

[1] ASL (Applied Science Laboratories), Eye Tracking System Instructions ASL Eye-Trac 6000 Pan/Tilt Optics, EyeTracPanTiltManual.pdf, ASL Version 1.04 01/17/2006.

[2] ASL (Applied Science Laboratories), Eye Tracking System Instruction Manual Model 501 Head Mounted Optics, ASL 2001.

[3] O. Christmann, N. Carbonell, Browsing through 3D representations of unstructured picture collections: and empirical study. In A. Celentano, P. Mussio (Eds.), *Proceedings of ACM Working Conference on Advanced Visual Interfaces*, Venezia, Italy, May 23-26, 2006, 369-372.

[4] C. Joergrnsen, *Image Retrieval*, The Scarecrow Press Inc., Lanham, Maryland and Oxford 2003.

[5] B. Kules, H. Hang, C. Plaisant, A. Rose, B. Shneiderman, Immediate Usability: a Case Study of Public Access design for a Community Photo Library. *Interacting with Computers* 16 (2004), 1171-1193.

[6] J. Kustanowitz, B. Shneiderman, Meaningful Presentations of Photo Libraries: Rationale and Applications of Bi-level Radial Quantum Layouts. In *Proc. 5th ACM/IEEE Joint Conference on Digital Libraries*, New York: ACM Press (2005), 188-196.

[7] W.M. Newman, M.G. Lamming, *Interactive System Design*. Addison-Wesley, Harlow, 1996.

[8] I. Wickelgren, Brain-Computer Interface Adds a New Dimension. *Science* 306 (2004), 1878-1879.

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-23

Packet Loss Concealment Algorithm for VoIP Transmission in Unreliable Networks

Artur JANICKI and Bartłomiej KSIĘŻAK
Institute of Telecommunications, Warsaw University of Technology
e-mail: ajanicki@tele.pw.edu.pl, bksiezak@op.pl

Abstract. In this chapter, the authors propose an algorithm for packet loss concealment (PLC) in transmission over IP-based networks with high packet loss rate. The algorithm is a sender-receiver-based extension of ANSI T1.521a Annex B PLC standard for G.711 voice codec. It consists in adding to a transmitted packet redundant parameters describing speech signal in another packet. Efficiency of the proposed algorithm was verified using subjective Absolute Category Rating (ACR) method and objective PESQ algorithm, and compared with original ANSI T1.521a Annex B standard. The intelligibility of speech was assessed using Semantically Unpredictable Sentences (SUS) tests. For high packet loss rates, all assessment methods proved superiority of the proposed algorithm over the original ANSI standard. The ACR tests showed that the proposed method can maintain speech quality above 3 in MOS scale even for packet loss rates of 20%-25%.

Keywords. VoIP, packet loss concealment, speech quality, intelligibility tests

Introduction

Voice over IP (VoIP) technology nowadays provides users with speech transmission quality similar (or higher) to channel-switching telephony, provided that it uses network with guaranteed Quality of Service (QoS). However, the technology uses best-effort networks such as majority of the Internet, or exposed to burst errors as in WLAN's. Such networks can be quite unreliable in the sense of packet loss and packet delays. If the network is highly congested, the delay values increase and voice packets do not arrive on time, resulting again in packet loss. In this case, the quality of the transmitted speech can be severely impaired.

There are different strategies how to cope with packet loss in VoIP transmission. They are known as packet loss concealment (PLC) algorithms and some of them are briefly presented in the next section.

In this chapter, the authors propose a PLC algorithm able to cope with high packet loss rates (PLR). On the receiver side it is an improved version of ANSI T1.521a Annex B algorithm elaborated for popular ITU-T G.711 codec (PCM 64 kbps), but it adds slight redundancy on the sender side, to improve the quality of transmitted speech in lossy IP-based network environment. Section 2 explains in details suggested algorithm. A comparative study between original ANSI T1.521a Annex B standard, the

proposed PLC algorithm and no-PLC version is then presented. Section 3 describes testing methodology and section 4 presents results of experiments and their discussion.

1. Packet Loss Concealment Schemes

PLC methods can be generally divided into coder-dependent and coder-independent techniques [16]. The former ones are codec-oriented – they use specificity of given coder operation and for example prevent them from loss of synchrony in case of packet loss. An example of PLC for G.722 codec is proposed in [13], where authors suggest adding site information in order to enable packet regeneration and to help the decoder to re-synchronize after a packet loss.

Coder-independent techniques do not need to know much about the used codec. Depending on where the PLC algorithm operates, they can be divided into:

- receiver-based techniques, such as ANSI T1.521a Annex B standard [1] developed for G.711 codec, which is reconstructing the lost packet(s) using parameters (linear prediction coefficients, F0, excitation) of the last correctly received packet. Another example using Hidden Markov Models (HMM) for estimating lost packet parameters is presented in [11];
- sender-receiver-based algorithms, using e.g. Forward Error Correction (FEC), retransmission schemes or adding redundancy.

Redundancy-based techniques can be quite effective as for quality improvement, but require additional bandwidth. The idea is that a voice packet carries ("piggybacks") redundant data for the other packets. In the simplest approach, it can carry duplicated neighbouring packet. Techniques that are more sophisticated consist in duplicating another packet, but encoded with a different voice codec, usually less bandwidth demanding. Such a method of combining G.711, ADPCM, GSM 06.10 is proposed in [3].

Other researchers proposed adding redundancy in the form of parameters of other packets: LPC (Linear Prediction Coding) coefficients [7], voicing and F0 (fundamental frequency) information [12], energy and zero crossing ratio [6]. In [14] the authors propose adding redundancy in the form of excitation parameters added to the preceding packet, but only for the most important packets. However, the lower PLR values are the main application area of the proposed methods.

2. Proposed Algorithm

This section presents the idea of the proposed algorithm for packet loss concealment. Its aim is to cope with high PLR values. On the receiver side, the algorithm is in fact using a modified ANSI T1.521a Annex B standard, which takes different parameters. In addition, the proposed solution involves changes on the sender side, so this technique becomes sender-receiver-based.

2.1. Drawbacks of ANSI T1.521a Annex B

ANSI T1.521a Annex B is a receiver-based PLC technique for G.711, which performs satisfactory if only single packets are lost. For unreliable networks with high PLR

values, when the loss of consecutive 2-3 or more packets is quite likely, the output speech quality is impaired, because the algorithm keeps reproducing the last correctly received packet many times. The only difference is the decreasing amplitude of the samples according to the scaling function. This causes a discomfort for a listener and impacts speech intelligibility, as it lengthens some phonemes and often allows other ones to disappear. It can be especially noticeable and harmful when speech is changing from voiced to unvoiced or vice versa.

2.2. Idea of the Proposed Algorithm

The proposed algorithm eliminates the described above problem by adding some redundancy to the encoded signal. To do this we suggest introducing the following changes at the sender side:

- implementing buffering of τ packets;
- adding to the currently transmitted n-th packet some parameters of the packet number $n+\tau$.

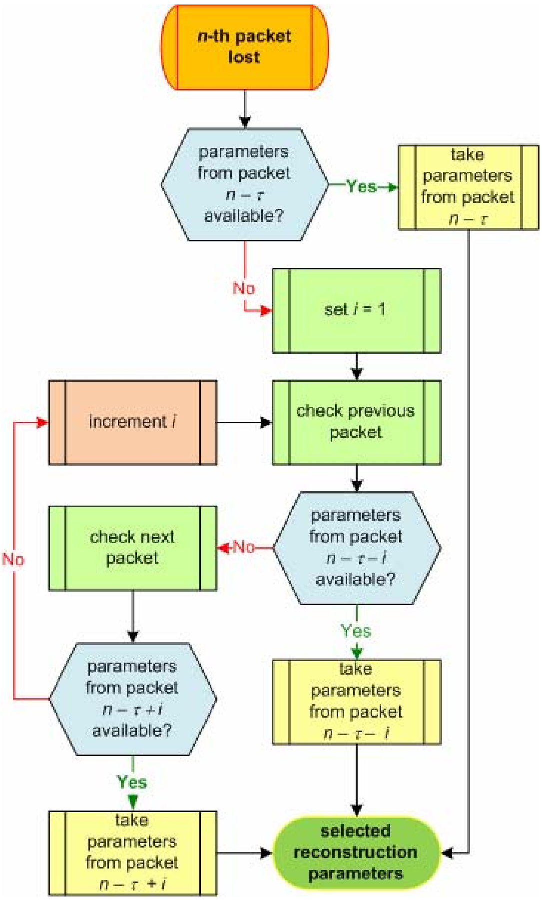


Figure 1. Algorithm for selection of reconstruction parameters

Following other research results mentioned in Section 1 it is proposed to select the parameters in order to characterize speech signal contained in a given packet in the best possible way and preserving a concise form. The method should be compliant with a G.711 decoder working with ANSI T1.521a Annex B PLC algorithm and therefore the following parameters were selected:

- 20 LPC coefficients, characterizing the speaker's vocal tract transmittance in a given time instant;
- *T0* parameter – a pitch period, equal 1/F0, where F0 is a pitch of a packet being characterized;
- *E* – short-time energy of the signal.

If a receiver realizes that the current packet is lost, it tries to reconstruct the packet. It uses the excitation signal of the last correctly received packet (as in ANSI T1.521a Annex B standard), but with LPC parameters taken from packet number n-τ and with *T0* value proper for the lost packet, instead of repeating *T0* of the last correctly received packet. The *E* parameter determines the scale the whole packet to keep the same level of energy of the reconstructed signal as in the original one.

If packet n-τ is not available (because it got lost, too), then the closest neighbouring set of parameters should be used to reconstruct the speech signal (see **Figure 1**).

2.3. Additional Requirements of the Proposed PLC Algorithm

The proposed algorithm has some drawbacks if compared with ANSI T1.521a Annex B. Firstly it requires changes on the transmitter side, so e.g. the transmitting softphone or a VoIP router has to be "PLC-algorithm aware". It also introduces additional computational load, because the signal parameters, unlike as in ANSI T1.521a Annex B standard, would need to be calculated (on a sender side) for every packet, not only in case of packet loss (on a receiver side).

The elaborated algorithm also forces the packets to carry increased payload, due to redundancy. The Last but not the least, it is introducing additional delay due to signal buffering in the encoder. The higher the τ parameter is, the higher the delay is. Too low τ can make the PLC algorithm not prove to losses of even small group of packets.

Anyway, the authors believe that potential benefits of this algorithm, when working in a network of high PLR, can overcome some drawbacks of this approach.

2.4. Implementation of the Proposed PLC Method

The proposed PLC algorithm was implemented in Matlab™ environment, using Voicebox toolbox for speech processing [15]. The LPC parameters were encoded by first converting them to reflection coefficients (RF), so that they ranged between (-1; 1) for easier quantization. A linear 7-bit quantizer was used for quantization of RF values, so the prediction coefficients filled up 18 bytes.

T0 (the pitch period in samples) occupied the next byte, ranging from 16 to 255, corresponding to F0 of 50-500 Hz. During experiments it turned out that the most precise F0 detection was achieved using *pitchdetat* algorithm [17], so this one was used. For unvoiced speech T0 value was chosen randomly, so that even if a unvoiced packet was reconstructed using excitation signal of a voiced one it was loosing its harmonicity.

E – energy parameter was encoded using a quantizer with logarithmic characteristics and stored as the last byte of total 20 redundant ones.

Assuming 20 ms frame size and sending one speech frame in one Ethernet packet, the increase of payload of a G.711 packet is not high: the redundant parameters cause 12.5% overhead, see **Figure 2**.

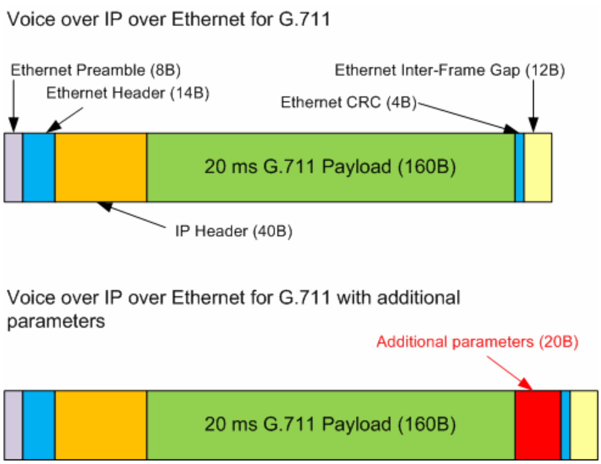


Figure 2. Content of Ethernet packets for original G.711 (above) and for G.711 with redundant parameters for PLC

The value of τ had to be chosen as a trade-off between efficiency and delay and experimentally set to 5. This is a quite high value, but lower values met in other works (e.g. 1-3 in [3]) turned to be insufficient for higher PLR.

Figure 3 presents visualization of packet loss concealment using both ANSI T1.521a Annex B and the proposed algorithm. Even by observing only the waveform one can notice that in case of severe signal damage (5 packets lost) the proposed method reconstructed the speech signal much better than the ANSI T1.521a Annex B standard, which kept copying the last properly received packet regardless of voicing loss, phoneme's boundary etc.

3. Testing Methodology

To verify if the proposed packet loss algorithm performs better than ANSI T1.521a Annex B standard, comparative experiments were run. In total 24 semantically unpredictable sentences (SUS) were recorded in Polish by a male and female speaker. They were transmitted through software packet network simulator at different packet loss rate (0-35%), using Gillbert-Elliott model of packet loss [8]. For a given PLR the same pattern of loss packets was maintained when testing all PLC methods. Three different PLC variants were used:

- no PLC algorithm at all;
- lost packets reconstructed using PLC algorithm according to ANSI T1.521a Annex B standard;
- lost packets reconstructed by the proposed PLC algorithm.

The reconstructed signal was recorded, thus forming 3 sets of output speech signal, henceforth called: no PLC, T1.521a, proposed PLC.

As quality assessment methods the following measures were selected:

- PESQ (Perceptual Evaluation of Speech Quality) algorithm for objective quality measure;
- ACR (Absolute Category Rating) method for subjective quality assessment;
- SUS (Semantically Unpredictable Sentences) tests for intelligibility testing.

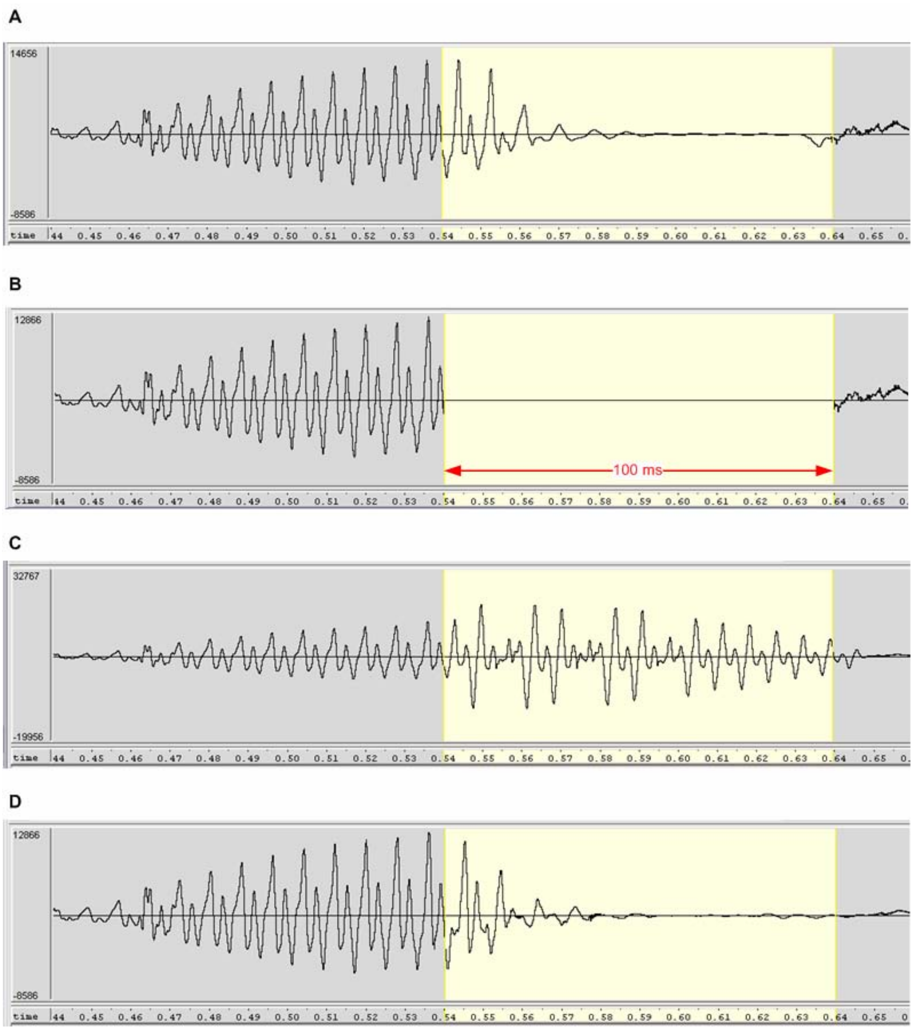


Figure 3. Illustration of packet reconstruction: (A) original signal, (B) signal with 5 packets loss and no PLC, (C) signal reconstructed with ANSI T1.521a Annex B, (D) signal reconstructed with the proposed algorithm

PESQ is an ITU-T standard for objective double-sided quality assessment for speech [9], which estimates listener's opinion by comparing transmitted and original speech signal, using human perception model.

ACR is a subjective method, which can be used for assessment of speech quality; among others, it can be used for testing VoIP transmission quality [4]. It requires

participation of listeners who assess the output speech quality, assigning scores on a 5-degree MOS (Mean Opinion Score) scale ranging from very bad to excellent.

SUS method of intelligibility testing consists in transmitting nonsense (semantically unpredictable) sentences through a telecommunication channel. A sample SUS sentence in English can be:

A table dances through the yellow future.

Listeners also need to participate in this test – they are asked to write down the sentence as they can hear it. Due to lack of sense in SUS sentences, they are able to correctly note them only if the speech is intelligible enough. This method has been previously successfully used in testing speech synthesis [2] and in VoIP intelligibility assessment [10].

PESQ and ACR methods result in MOS score (0.5 - 4.5), while intelligibility tests result in percentage of correctly conveyed sentences and words.

48 subjects took part in SUS and ACR tests; they were exposed to 576 recordings (24 SUS sentences x 8 PLR values x 3 PLC algorithms), so that each version was assessed by 2 listeners. The results are presented and discussed in the next section.

4. Results of Experiments

Figure 4 presents average MOS results achieved both from calculations of PESQ algorithm (objective method) and from ACR technique during listening tests (subjective method).

Both tests showed obviously the maximum score (4.5 MOS) for signal with no loss. They also demonstrated decreasing quality along with increasing packet loss rate and none of algorithms was able to prevent it. However, in both tests the proposed PLC algorithm significantly outperformed the original ANSI T1.521a Annex B standard. According to subjective assessment transmission over a network with PLR equal 25% resulted still in MOS above 3. The higher packet loss was, the higher MOS gain was observed, sometimes exceeding even 0.5 MOS.

What is more, it turned out that for heavily unreliable (lossy) networks a receiver with ANSI T1.521a Annex B standard PLC algorithm according to subjective judgments in ACR tests performs even worse than lack of any PLC algorithm at all. Apparently, the listeners preferred silence in lost frames rather than several times repeated last correctly received packet, as it is according to the ANSI standard.

The results of both subjective and objective methods were highly correlated – the correlation coefficient between them was 0.9414, what confirms results reliability. The results for higher PLR values are better for ACR than for PESQ – probably if the listeners were able to understand the utterance, then they also were more favourably disposed when assigning the MOS value. PESQ algorithm proved to be more demanding.

Figure 5 presents an example of MOS variation in PESQ test for individual SUS sentences. It shows that some sentences were less damaged by the loss of packets and got higher MOS results than the others get. In addition, the figure indicates that the proposed PLC helped almost in all the cases. The only exception was the sentence number 15, where the gain was hardly noticeable.

Figure 6 and **Figure 7** present results of intelligibility tests, showing respectively word and sentence intelligibility. Here also the results worsen when packet loss ratio is increasing. Absolute sentence intelligibility is lower than the word one, because a SUS sentence is disqualified even if only one constituting word was wrongly received.

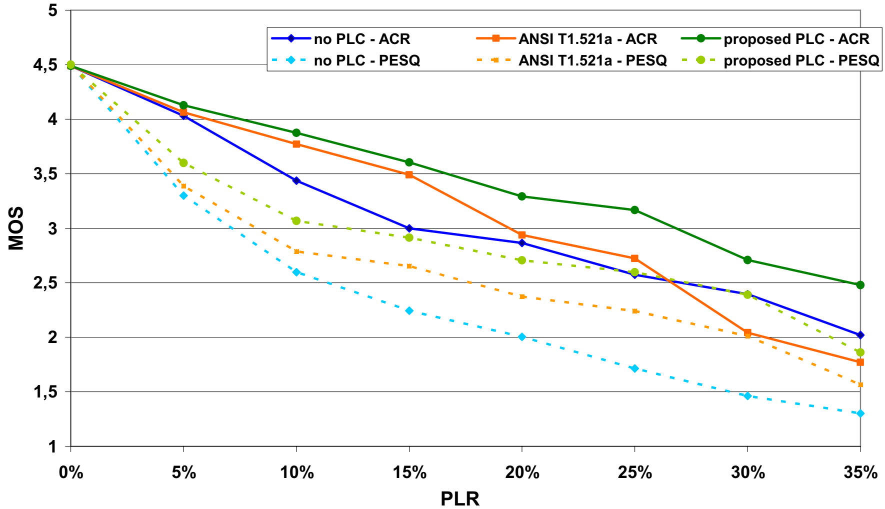


Figure 4. MOS results for different PLR values using subjective ACR method and objective PESQ algorithm

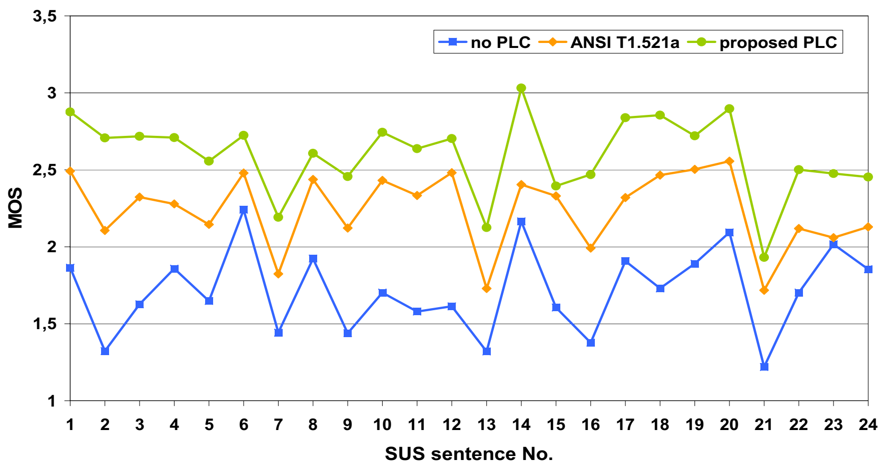


Figure 5. Sample MOS results using PESQ algorithm for individual SUS sentences (PLR = 25%)

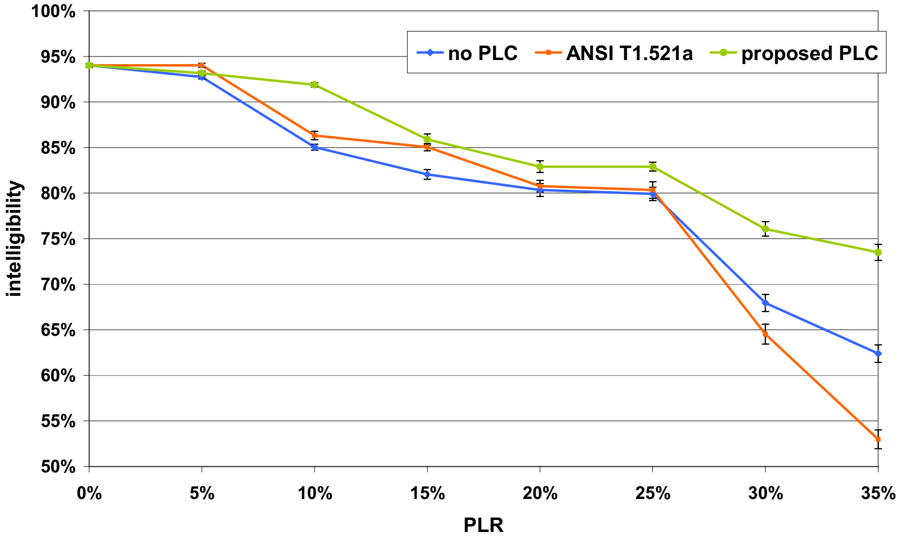


Figure 6. Word intelligibility results from SUS tests

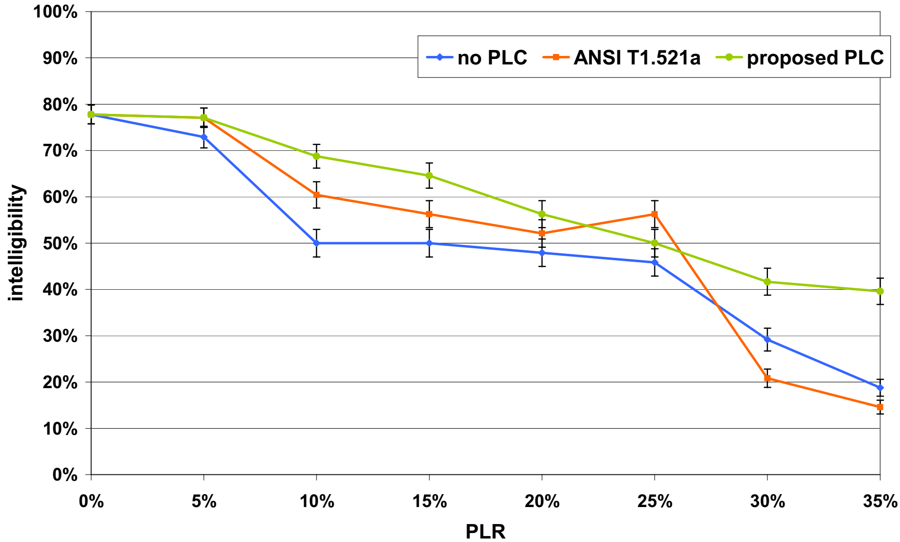


Figure 7. Sentence intelligibility results from SUS tests

Results of intelligibility confirm the conclusions from the speech quality tests. While the intelligibility results are very similar or identical for 0% and 5% PLR values, the proposed PLC algorithm performs significantly better than ANSI T1.521a Annex B standard for almost all PLR values higher than 5%. For the rate of 35%, the difference exceeds even 10% and 20%, respectively for word and sentence intelligibility.

A loss rate of 25% is a remarkable exception however – in this case, sentence intelligibility for the original PLC algorithm is more than 10% higher than for the proposed one, and higher than for original algorithm at 20% packet loss rate. This local increase of sentence intelligibility was confirmed neither by word intelligibility result, nor by MOS scores. Closer analysis of this particular case showed that accidentally the wrongly heard words happened to be grouped together in the same SUS sentences, while the other SUS sentences were relatively intelligible, thus causing higher sentence intelligibility score.

The confidence intervals marked in the intelligibility figures, estimated for confidence level of 0.9, prove satisfactory trustworthiness of the presented results.

5. Conclusions

The results of quality and intelligibility tests showed that the proposed PLC algorithm improves significantly the quality of speech transmitted over an unreliable network with a high packet loss rate. It performs also better than a receiver-based ANSI T1.521a Annex B algorithm; however, it introduces additional delay, which has to be taken into account. Anyway, an increased delay is often a necessary expense whenever the signal quality is a priority.

The proposed PLC algorithm is in fact an extension of ANSI T.521a Annex B to a sender-receiver-based version and it is compliant with its original version – the receiver can switch to work according to ANSI T1.521a Annex B standard if it finds no redundant data in a G.711 packet.

τ parameter is adjustable – it can be modified, also dynamically depending on the network conditions: it can be decreased e.g. if the WLAN link quality improves. An optional additional byte (or part of it, as 3 bits are enough) can be set to inform about the current value of τ parameter.

References

[1] ANSI American National Standard, *Supplement to T1.521-1999, Packet Loss Concealment for Use with ITU-T Recommendation G.711*, Alliance for Telecommunications Industry Solutions, Washington, 2000.

[2] Ch. Benoit, M. Grice, and V. Hazan, The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences, *Speech Communication 18*, Elsevier Science B.V., 1996, pp. 381-392.

[3] J. C. Bolot, and A. V. Garcia: Control mechanisms for packet audio in the Internet, *In Proc. of IEEE INFOCOM*, Apr. 1996, pp. 232-239.

[4] S. Brachmański: VoIP – quality assessment of speech transmission using ACR and DCR methods (in Polish), *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne* 8-9/2003, Warsaw, 2003.

[5] G. Carle, T. Hoshi, L. Senneck, and H. Le: Active Concealment for Internet speech transmission, *International Workshops on Active Networks (IWAN),* Tokyo, 2000.

[6] N. Erdol, C. Castelluccia, and A. Zilouchian: Recovery of missing speech packets using the short-time energy and zero-crossing measurements, *IEEE Trans. on Speech and Audio Processing*, 1(3), July 1993, pp. 295-303.

[7] V. Hardman, M. A. Sasse, M. Handley, and A. Watson: Reliable audio for use over the Internet, In *Proc. of Int'l Networking Conf.*, June 1995, pp. 171-178.

[8] ITU-T COM 12-D104-E: *Modelling Burst Packet Loss within the E Model*, Geneva, 27-31 January 2003.

[9] ITU-T *Recomendation P.862 Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codec*, Geneva, 02.2001.
[10] A. Janicki, B. Księżak, J. Kijewski, and S. Kula: Speech quality assessment in the Internet telephony using semantically unpredictable sentences (in Polish), *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne* 8-9/2006, Warsaw, 2006.
[11] Ch. A. Rødbro, M. N. Murthi, S. V. Andersen, and S. H. Jensen: Hidden Markov Model-Based Packet Loss Concealment for Voice Over IP, *IEEE Transactions On Audio, Speech, And Language Processing*, Vol. 14, No. 5, 10. 2006
[12] H. Sanneck: Concealment of lost speech packets using adaptive packetization, In IEEE Int'l Conf. on Multimedia Computing and Systems, 1998, pp. 140–149.
[13] N. Shetty, and J, D. Gibson: Packet Loss Concealment for G.722 using Side Information with Application to Voice over Wireless LANs, *Journal of Multimedia*, vol. 2, No. 3, June 2007, pp. 66-76.
[14] L. Tosun, and P. Kabal, Dynamically Adding Redundancy for Improved Error Concealment in Packet Voice Coding, In *Proc. European Signal Processing Conf.* (Antalya, Turkey), 4 pp. , Sept. 2005
[15] Voicebox: Speech Processing Toolbox for MATLAB, *http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html*
[16] B. W. Wah, X. Su, and D. Lin: A survey of error-concealment schemes for real-time audio and video transmissions over the internet, In *Proc. Int'l Symposium on Multimedia Software Engineering*, Taipei, Taiwan, December 2000, pp. 17-24.
[17] A. Wong "Pitchdetat", 01.04.2003 *http://hebb.mit.edu/courses/9.29/2003/athena/auditory/birdsong/pitchdetat.m*

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-34

Effectiveness of Video Segmentation Techniques for Different Categories of Videos

Kazimierz CHOROŚ [a] and Michał GONET [b]
Wrocław University of Technology, Institute of Applied Informatics,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
[a] *e-mail: choros@pwr.wroc.pl*
[b] *e-mail: mgonet@vp.pl*

Abstract. Visual retrieval systems as well as Internet search engines demand efficient techniques of indexing to facilitate fast access to the required video or to the required video sequences in the video databases. Digital video databases are more and more frequently implemented not only in the Internet network but also in local networks and even in local personal computers. Different approaches to digital video indexing are applied: textual approach, extraction and representation of technical or structural features, content-based analysis, and finally segmentation. The segmentation process leads to the partition of a given video into a set of meaningful and individually manageable segments, which then can serve as basic units for indexing. Video has temporal properties such as camera motion, object movements on the scene, sequential composition, and interframe relationships. An effective segmentation technique is able to detect not only abrupt changes but also gradual scene changes, such as fade and dissolve transitions. The nature of movies, mainly the style of video editing has an influence on the effectiveness of temporal segmentation methods. The effectiveness of four methods was analyzed for five different categories of movie: TV talk-show, documentary movie, animal video, action & adventure, and pop music video. The cuts have been recognized as well as cross dissolve effects. The tests have shown that the specific nature of videos has an important influence on the effectiveness of temporal segmentation methods.

Keywords. digital video indexing, temporal segmentation, scene changes, cuts, cross dissolve effects, movie categories

Introduction

Multimedia data are very popular in local retrieval systems as well as in the Internet network. The Internet search engines demand efficient techniques of indexing not only of textual parts of Internet pages but also of images, audio and video clips. The text is auto descriptive. The text is composed of natural language words which can be treated as index terms used in retrieval. The process of indexing is usually limited to the identification of words or expressions in the text.

In many cases the textual information retrieval systems also retrieve other multimedia objects. They are processed similarly to texts, i.e. the multimedia are indexed by textual indexes derived from the texts accompanying these multimedia

objects, for example from the text of captions of the images. A real multimedia information retrieval system should analyze the information content in all media forms and the user should be able to express in his request different aspects of the media he needs. This leads to the necessity of analyzing and of indexing the multimedia documents using specific features of each media. The other media are not of the same nature as text. Images as well as audio and video clips are much more expressive than traditional documents containing only text but they do not explicitly include index terms. The multimedia information stored in a document should be represented by a set of specific features, i.e. by textual, structural as well as by the content-descriptive metadata specific for each medium [5], [6], [7].

The indexing process of digital video clips can be performed in three levels:

- text annotation – traditional approach leading to the identification of index terms describing not only the content of the video clip but also technical features [5], [19];
- content analysis – automatic analysis of video content [1], [8];
- temporal segmentation – leading to the identification of meaningful video units, like shots, episodes or scenes [4], [11], [12], [14], [16], [17], [18].

The text annotation process is based on traditional indexing process but in case of multimedia data the indexing methods should take into account the specific features of multimedia. All visual information can be described by traditional bibliographic data such as title, abstract, subject, genre, also by technical or structural data such as colour, texture, shape for a static image, or a hierarchy of movie, segment, motion, pan, zoom, scene, and finally shot for a movie [2], [5], [7], [9], [13]. These characteristics are derived on the bases of textual description of a medium or from relatively easy measured technical and structural parameters. Therefore, they are called content-independent metadata. Text could be also present in a video [10]. Text, like title or names of movie stars and other artists or the name of the director, is usually superimposed on the images, or included as closed captions.

Let's notice that the text has also structural features. These are for example the language of the text, its length, number of standard pages or paragraphs, number of words or characters, number of columns, the size of characters, and many other statistical and editorial properties.

Undoubtedly users could benefit from a retrieval in which one could refer to content semantics, i.e. to events, emotions or meaning associated with visual information. These data are called content-descriptive metadata. Content of a still image [7] is expressed by its perceptual properties, like colour, texture, shape, and spatial relationships, but also by semantic primitives corresponding to abstractions, like objects, roles, and scenes, and finally by such imprecise features like impressions, emotions, and meaning associated with the perceptual features combined together.

A video clip is also structured into a strict hierarchy. A clip is divided into some scenes, and a scene is composed of one or more camera shots. A scene usually corresponds to some logical event in a video such as a sequence of shots making up a dialogue scene, or an action scene in a movie. The temporal segmentation of video clips is a process of detection of the shot changes present in the video sequences. A shot is usually defined as a continuous video acquisition with the same camera, so, it is as a sequence of interrelated consecutive frames recorded contiguously and representing a continuous action in time or space.

A shot change occurs when a video acquisition is done with another camera. The cut is the simplest and the most frequent way to perform a change between two shots.

In this case, the last frame of the first video sequence is directly followed by the first frame of the second video sequence. Cuts are probably the easiest shot changes to be detected. But software used for digital movie editing is more and more complex and other shot changes are now available. They include effects or transitions like a wipe, a fade, or a dissolve [15]. A wipe effect is obtained by progressively replacing the old image by the new one, using a spatial basis. A dissolve is a transition where all the images inserted between the two video sequences contain pixels whose values are computed as linear combination of the final frame of the first video sequence and the initial frame of the second video sequence. Cross dissolve describes the cross fading of two scenes. Over a certain period of time (usually several frames or several seconds) the images of two scenes overlay, and then the current scene dissolves into a new one. Fades are special cases of dissolve effects, where a most frequently black frame replaces the last frame of the first shot (fade in) or the first frame of the second shot (fade out). There are also many other kinds of effects combining for example wipe and zoom, etc.

1. Temporal Segmentation

A temporal segmentation is a process of partitioning a video sequence into shots and scenes. It is the first and indispensable step towards video-content analysis and content-based video browsing and retrieval.

There are many methods of temporal segmentation and they are still being refined [3], [19], [22]. The simplest methods are those analyzing individual pixels of the consecutive frames, the most complex are based on the histogram analysis and the detection of the motion during the video sequence. Several categories can be distinguished [12]:

- pixel pair differences,
- histogram differences of frames,
- block sampling of video frames,
- feature-based methods,
- motion-based methods, and
- combination of approaches.

The fundamental method is a shot change detection method based on pixel pair difference. It is known as the template matching. For every two successive frames, the values of intensity differences are computed for pixel pairs having the same spatial position in these two frames. Then, the cumulated sum of intensity differences is compared with a fixed threshold in order to determine if a shot change has been detected. Several other statistical measures have been proposed, among others the normalized difference energy and the normalized sum of absolute differences.

Two images can be also compared using global features instead of local features (pixels). Histogram is the best global image feature widely used in image processing. The main advantage of histogram-based methods is their global aspect. There are several methods which compute differences between histograms or weighted differences between histograms. The other methods define an intersection operator between histograms, different distances or different similarity measures.

The main advantage of the third group of segmentation methods that use the block sampling of video frames is the decrease of the computation time. The algorithms

proposed for pixel or histogram analysis methods are applied for block representation of images. The other advantage of block-based methods is their relatively weak sensitivity to noise and camera or object motion.

The next group of temporal segmentation methods takes into consideration sophisticated features of images, such as [12]:

- moment invariants combined with histogram intersection,
- contour lines extracted from the image,
- some feature points extracted using Hough transform,
- planar points – points contained in a flat neighbourhood of the hyper surface,
- comparing of colour histograms of regions of two successive frames,
- modelling of the video transition effects,
- use of some decision process as Bayesian methods,
- features computed from classical statistical approaches, and
- use of hidden Markov models.

Among the motion-based methods are the methods based on global (or camera) motion, motion vectors, optical flow, and correlation in the frequency domain.

The last group of methods relays upon the shot change detection by combining two or several algorithms.

2. Video Features and Categories

The question arises whether the dynamic nature and editing style of video clips has an influence on the effectiveness of temporal segmentation methods. Generally, the more dynamic video editing, the easier cut detection. In a dynamic clip there are usually no cross dissolve effects.

Most frequently categories of movies are defined on the basis of theme of movies. But we can define other criteria to categorize movies. The dynamism, style, and richness of movie storyboard and editing are very significant:

- dialog (audio) speed or pace of the narration,
- camera motion,
- light changes,
- action dynamism, and
- special effects.

We have decided to measure the influence of the movie features on effectiveness of the temporal segmentation using five digital videos of different types: TV talk-show, documentary movie, animal video, action & adventure, and pop music video.

TV talk-show is generally special video shot realized in the TV studio, with a static scene, almost without camera movements and without object movements, without colour changes of the scene, without light effects and without special editing effects.

Documentary video is also static in nature, also without dynamic changes on the scene, but some fragments of such a video could be clearly different. In documentary videos many effects such as fades and dissolves are usually used.

Animal videos are totally different. Objects of the scene (mainly animals) are moving, but rather in a slaw manner, also camera is moving also in a slow manner, rather constant but very rich variety of colour used in every episode. The dissolve effects are rather long.

Adventure video is a video composed of dynamic, short clips. We observe in such movies dynamic camera movements, also dynamic object movements on the scene, rich colours, contrast quickly changing, changing light, and many effects.

In pop music videos the editing style is extremely dynamic, many images are rapidly juxtaposed together, many very rapid movements of cameras, lights, as well as objects of the scene. The clips of the video are very short, even less then 1 second, cuts are very frequent, dissolve effects are relatively rare, contrary to light effects which are very often used.

The following five representatives of the discussed above types of videos were used in experiments with temporal segmentation methods:

- **TV talk-show** – resolution: 576x432, duration: 5:05, frame rate: 25 fps, number of cuts: 61, number of cross-dissolves: 0;
- **documentary video** – resolution: 512x320, duration: 5:12, frame rate: 25 fps, number of cuts: 45, number of cross-dissolves: 8;
- **animal video** – resolution: 640x352, duration: 5:01, frame rate: 25 fps, number of cuts: 66, number of cross-dissolves: 2;
- **adventure video** – resolution: 640x272, duration: 5:06, frame rate: 25 fps, number of cuts: 98, number of cross-dissolves: 0;
- **POP music video** – resolution: 436x4370, duration: 1:56, frame rate: 25 fps, number of cuts: 93, number of cross-dissolves: 13.

3. Tests and Results

Four methods of temporal segmentation have been implemented. These methods are:

- pixel pair differences [7], [21],
- likelihood ratio method based on block sampling of video frames [2], [9],
- histogram differences of frames [7], [13],
- twin threshold comparison [22].

These methods process non-compressed video files. Experimental results will be presented to demonstrate the performance of various algorithms. The algorithms have some parameters. In the tests several sets of specific parameters have been used to demonstrate their influence on the effectiveness of the temporal segmentation methods.

The Table 1 to 4 presents the results obtained in the tests.

Notations used in the Tables:

PT - processing time,
C - number of recognized cuts,
CC - number of cuts correctly recognized,
OC - number of omitted cuts (not-recognized),
FC - faulty cuts,
D - number of recognized cross-dissolve effects,
CD - number of cross-dissolves correctly recognized,
OD - number of omitted cross-dissolves (not-recognized),
FD - faulty cross-dissolves,
R - recall,
P - precision.

Table 1. Experimental results of temporal segmentation for pixel pair differences.

Two pixels are different if the difference of their values is greater than t.
There is a cut between two frames if T % of pixels of two consecutive frames is different.

Para-meters			Cuts				Cross-Dissolves				Effectiveness	
TV Talk-Show												
t	**T**	**PT**	**C**	**CC**	**OC**	**FC**	**D**	**CD**	**OD**	**FD**	**R**	**P**
20	20	07:40	278	61	0	217	0	0	0	0	**1.00**	**0.22**
30	50	07:45	61	61	0	0	0	0	0	0	**1.00**	**1.00**
40	30	07:48	61	61	0	0	0	0	0	0	**1.00**	**1.00**
100	20	07:34	17	17	44	0	0	0	0	0	**0.28**	**1.00**
Documentary Video												
t	**T**	**PT**	**C**	**CC**	**OC**	**FC**	**D**	**CD**	**OD**	**FD**	**R**	**P**
20	20	05:21	524	45	0	479	0	1	7	0	**0.87**	**0.09**
30	50	05:24	43	43	2	0	0	0	8	0	**0.81**	**1.00**
40	30	05:20	45	45	0	0	0	0	8	0	**0.85**	**1.00**
100	20	05:19	36	36	9	0	0	0	8	0	**0.68**	**1.00**
Animal Video												
t	**T**	**PT**	**C**	**CC**	**OC**	**FC**	**D**	**CD**	**OD**	**FD**	**R**	**P**
20	20	06:45	925	60	6	865	0	0	2	0	**0.88**	**0.06**
30	50	06:48	53	52	14	1	0	0	2	0	**0.76**	**0.98**
40	30	06:48	74	54	12	20	0	0	2	0	**0.79**	**0.73**
100	20	06:50	22	22	44	0	0	0	2	0	**0.32**	**1.00**
Adventure Video												
t	**T**	**PT**	**C**	**CC**	**OC**	**FC**	**D**	**CD**	**OD**	**FD**	**R**	**P**
20	20	05:37	2596	98	0	2498	0	0	0	0	**1.00**	**0.04**
30	50	05:37	132	91	7	41	0	0	0	0	**0.93**	**0.69**
40	30	05:36	274	97	1	177	0	0	0	0	**0.99**	**0.35**
100	20	05:37	65	52	46	13	0	0	0	0	**0.53**	**0.80**
POP Music Video												
t	**T**	**PT**	**C**	**CC**	**OC**	**FC**	**D**	**CD**	**OD**	**FD**	**R**	**P**
20	20	03:07	585	95	0	490	0	8	5	0	**0.95**	**0.18**
30	50	03:09	72	70	25	2	0	2	11	0	**0.67**	**1.00**
40	30	03:10	105	95	0	10	0	2	11	0	**0.90**	**0.92**
100	20	03:03	57	57	38	0	0	0	13	0	**0.53**	**1.00**

Table 2. Experimental results of temporal segmentation for the likelihood ratio method based on block sampling of video frames.

Two blocks are different if the difference of their average values is greater than t. There is a cut between two frames if T % of blocks of two consecutive frames is different.

Parameters				**Cuts**				**Cross-Dissolves**				**Effectiveness**	
TV Talk-Show													
blocks	**t**	**T**	**PT**	**C**	**CC**	**OC**	**FC**	**D**	**CD**	**OD**	**FD**	**R**	**P**
4x3	2	30	14:48	51	50	11	1	0	0	0	0	**0.82**	**0.98**
4x3	3	20	14:46	44	43	18	1	0	0	0	0	**0.70**	**0.98**
8x6	2	40	14:36	68	61	0	7	0	0	0	0	**1.00**	**0.90**
16x9	1	80	14:48	296	61	0	235	0	0	0	0	**1.00**	**0.21**
Documentary Video													
blocks	**t**	**T**	**PT**	**C**	**CC**	**OC**	**FC**	**D**	**CD**	**OD**	**FD**	**R**	**P**
4x3	2	30	09:39	41	41	4	0	0	0	8	0	**0.77**	**1.00**
4x3	3	20	09:32	40	40	5	0	0	0	8	0	**0.75**	**1.00**
8x6	2	40	09:45	45	45	0	0	0	0	8	0	**0.85**	**1.00**
16x9	1	80	09:46	638	45	0	593	0	7	1	0	**0.98**	**0.08**
Animal Video													
blocks	**t**	**T**	**PT**	**C**	**CC**	**OC**	**FC**	**D**	**CD**	**OD**	**FD**	**R**	**P**
4x3	2	30	12:15	67	60	6	7	0	0	2	0	**0.88**	**0.90**
4x3	3	20	12:23	73	58	8	15	0	0	2	0	**0.85**	**0.79**
8x6	2	40	12:08	82	65	1	17	0	0	2	0	**0.96**	**0.79**
16x9	1	80	12:16	1587	66	0	1521	0	2	0	0	**1.00**	**0.04**
Adventure Video													
blocks	**t**	**T**	**PT**	**C**	**CC**	**OC**	**FC**	**D**	**CD**	**OD**	**FD**	**R**	**P**
4x3	2	30	09:57	118	86	12	32	0	0	0	0	**0.88**	**0.73**
4x3	3	20	09:59	100	76	22	24	0	0	0	0	**0.78**	**0.76**
8x6	2	40	09:54	142	93	3	47	0	0	0	0	**0.95**	**0.65**
16x9	1	80	09:55	2839	98	0	2741	0	0	0	0	**1.00**	**0.03**
POP Music Video													
blocks	**t**	**T**	**PT**	**C**	**CC**	**OC**	**FC**	**D**	**CD**	**OD**	**FD**	**R**	**P**
4x3	2	30	06:04	72	60	35	12	7	4	9	3	**0.59**	**0.81**
4x3	3	20	06:06	65	52	43	13	0	5	8	0	**0.53**	**0.88**
8x6	2	40	06:03	102	87	8	15	0	5	8	0	**0.85**	**0.90**
16x9	1	80	06:04	0	0	95	0	0	0	13	0	**0.00**	**0.00**

Table 3. Experimental results of temporal segmentation for histogram differences of frames. There is a cut between two frames if T % of values of the histograms of two consecutive frames is different.

Parameter		**Cuts**				**Cross-Dissolves**				**Effectiveness**	
TV Talk-Show											
T	**PT**	**C**	**CC**	**OC**	**FC**	**D**	**CD**	**OD**	**FD**	**R**	**P**
20	05:36	61	61	0	0	0	0	0	0	**1.00**	**1.00**
40	05:40	39	39	22	0	0	0	0	0	**0.64**	**1.00**
60	06:19	0	0	61	0	0	0	0	0	**0.00**	**0.00**
80	06:18	0	0	61	0	0	0	0	0	**0.00**	**0.00**
Documentary Video											
T	**PT**	**C**	**CC**	**OC**	**FC**	**D**	**CD**	**OD**	**FD**	**R**	**P**
20	04:07	60	44	1	16	0	3	5	0	**0.89**	**0.78**
40	04:08	34	34	11	0	0	0	8	0	**0.64**	**1.00**
60	04:07	16	16	29	0	0	0	8	0	**0.30**	**1.00**
80	04:09	3	3	42	0	0	0	8	0	**0.06**	**1.00**
Animal Video											
T	**PT**	**C**	**CC**	**OC**	**FC**	**D**	**CD**	**OD**	**FD**	**R**	**P**
20	04:52	98	64	2	34	0	1	1	0	**0.96**	**0.66**
40	04:49	36	36	30	0	0	0	2	0	**0.53**	**1.00**
60	04:50	21	21	45	0	0	0	2	0	**0.31**	**1.00**
80	04:50	5	5	61	0	0	0	2	0	**0.07**	**1.00**
Adventure Video											
T	**PT**	**C**	**CC**	**OC**	**FC**	**D**	**CD**	**OD**	**FD**	**R**	**P**
20	04:16	135	90	8	45	0	0	0	0	**0.92**	**0.67**
40	04:15	61	48	50	13	0	0	0	0	**0.49**	**0.79**
60	04:15	22	19	79	3	0	0	0	0	**0.19**	**0.86**
80	04:17	5	5	93	0	0	0	0	0	**0.05**	**1.00**
POP Music Video											
T	**PT**	**C**	**CC**	**OC**	**FC**	**D**	**CD**	**OD**	**FD**	**R**	**P**
20	02:18	95	66	29	29	0	4	9	0	**0.65**	**0.74**
40	02:21	6	5	90	1	0	0	13	0	**0.05**	**0.83**
60	02:21	1	1	94	0	0	0	13	0	**0.01**	**1.00**
80	02:20	0	0	95	0	0	0	13	0	**0.00**	**0.00**

Table 4. Experimental results of temporal segmentation for the twin threshold comparison method.

In the twin comparison threshold method the twin thresholds Tb is used for cut detection and Ts is used for special effect detection.
l, k – additional parameters.

Parameter					Cuts				Cross-Dissolves				Effectiveness	
TV Talk-Show														
l	**k**	**T_s**	**T_b**	**PT**	**C**	**CC**	**OC**	**FC**	**D**	**CD**	**OD**	**FD**	**R**	**P**
1	2	2	9	12:55	85	61	0	24	78	0	0	78	**1.00**	**0.37**
2	4	3	17	13:02	61	61	0	0	32	0	0	32	**1.00**	**0.66**
3	6	5	24	12:49	59	59	2	0	14	0	0	14	**0.97**	**0.81**
4	8	6	32	12:38	55	55	6	0	7	0	0	7	**0.90**	**0.89**
Documentary Video														
l	**k**	**T_s**	**T_b**	**PT**	**C**	**CC**	**OC**	**FC**	**D**	**CD**	**OD**	**FD**	**R**	**P**
1	2	2	11	08:48	128	45	0	83	145	8	0	137	**1.00**	**0.19**
2	4	5	20	08:47	56	44	1	12	47	8	0	39	**0.98**	**0.50**
3	6	7	29	08:50	42	39	6	3	22	6	2	16	**0.85**	**0.70**
4	8	9	38	08:52	34	33	12	1	10	5	3	5	**0.72**	**0.86**
Animal Video														
l	**k**	**T_s**	**T_b**	**PT**	**C**	**CC**	**OC**	**FC**	**D**	**CD**	**OD**	**FD**	**R**	**P**
1	2	3	14	10:50	169	66	0	103	104	2	0	102	**1.00**	**0.25**
2	4	5	25	11:00	67	58	8	9	51	1	1	50	**0.87**	**0.50**
3	6	8	36	11:17	42	41	25	1	23	1	1	22	**0.62**	**0.65**
4	8	11	47	11:29	30	30	36	0	11	1	1	10	**0.46**	**0.76**
Adventure Video														
l	**k**	**T_s**	**T_b**	**PT**	**C**	**CC**	**OC**	**FC**	**D**	**CD**	**OD**	**FD**	**R**	**P**
1	2	4	16	09:33	154	95	3	59	167	0	0	167	**0.97**	**0.30**
2	4	7	28	09:35	99	76	22	23	38	0	0	38	**0.78**	**0.56**
3	6	11	40	09:37	59	47	51	12	7	0	0	7	**0.48**	**0.71**
4	8	15	52	09:38	36	30	68	6	4	0	0	4	**0.31**	**0.75**
POP Music Video														
l	**k**	**T_s**	**T_b**	**PT**	**C**	**CC**	**OC**	**FC**	**D**	**CD**	**OD**	**FD**	**R**	**P**
1	2	4	15	05:19	140	83	12	57	64	12	1	52	**0.88**	**0.47**
2	4	8	26	05:15	37	29	66	8	24	9	4	15	**0.35**	**0.62**
3	6	12	36	05:14	7	5	90	2	6	6	7	0	**0.10**	**0.85**
4	8	17	47	05:16	3	3	92	0	1	0	13	1	**0.03**	**0.75**

4. Best Results of Recall and Precision

Table 5. The best results of recall of temporal segmentation methods received for every category of video.

Results with the best RECALL	Pixel pair differences		Likelihood ratio method		Histogram differences		Twin threshold comparison	
	R	P	R	P	R	P	R	P
TV Talk-Show	1.00	1.00	1.00	0.89	1.00	1.00	1.00	0.66
Documentary Video	0.87	0.09	0.98	0.08	0.89	0.78	1.00	0.19
Animal Video	0.88	0.06	1.00	0.04	0.96	0.66	1.00	0.25
Adventure Video	1.00	0.04	1.00	0.03	0.92	0.67	0.97	0.30
POP Music Video	0.95	0.18	0.85	0.85	0.65	0.74	0.88	0.47

The highest values of recall have been obtained using the twin threshold comparison method, but for this method with the best parameters of processing and in consequence the best results of recall we observe relatively very low results of precision. The simple histogram difference method produced good results of recall as well as of precision.

Table 6. The best results of precision of temporal segmentation methods received for every category of video.

Results with the best PRECISION	Pixel pair differences		Likelihood ratio method		Histogram differences		Twin threshold comparison	
	R	P	R	P	R	P	R	P
TV Talk-Show	1.00	1.00	0.82	0.98	1.00	1.00	0.90	0.89
Documentary Video	0.85	1.00	0.85	1.00	0.64	1.00	0.72	0.86
Animal Video	0.32	1.00	0.90	0.89	0.53	1.00	0.46	0.76
Adventure Video	0.53	0.80	0.78	0.76	0.05	1.00	0.31	0.75
POP Music Video	0.67	1.00	0.85	0.90	0.01	1.00	0.10	0.85

In the case of precision the highest values of recall have been obtained using the histogram difference method. Notice, however, that the more dynamic editing style of

video, the lower the value of recall. For the adventure video and music video the value of recall is not acceptable (only 0.05 and 0.01)

Generally, when we like to receive the best values of the recall ratio, the twin threshold comparison method is recommended. The histogram difference method leads to the best values of the precision ration.

5. Conclusions

The segmentation process leads to the partition of a given video into a set of meaningful and individually manageable segments, which then can serve as basic units for indexing. Video has temporal properties such as camera motion, object movements on the scene, sequential composition, and interframe relationships. An effective segmentation technique is able to detect not only abrupt changes but also gradual scene changes, such as fade and dissolve transitions.

The nature of movies, mainly the style of video editing has an influence on the effectiveness of temporal segmentation methods. In the experiments and tests performed the effectiveness of four methods was analyzed for five different categories of movie: TV talk-show, documentary movie, animal video, action & adventure, and pop music video. The cuts have been recognized as well as cross dissolve effects.

The tests have shown that the specific nature of videos has an important influence on the effectiveness of temporal segmentation methods, fundamental methods of video indexing.

References

[1] M. Bertini, A. Del Bimbo, and P. Pala, Content-based indexing and retrieval of TV news, Pattern Recognition Letters 22, (2001), 503-516.

[2] R. Brunelli, O. Mich, and C.M. Modena, A Survey on the Automatic Indexing of Video Data, Journal of Visual Communication and Image Representation 10 (1999).

[3] L-H. Chena, Y-C. Laib, and H-Y.M. Liaoc, Movie scene segmentation using background information, Pattern Recognition 41 (2008), 1056-1065.

[4] L.-F. Cheong, Scene-based shot change detection and comparative evaluation, Computer Vision and Image Understanding 79 (2000), 224-235.

[5] K. Choroś, Retrieval criteria in visual information retrieval systems. In: Multimedia and Network Information Systems, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, 2006, 291-303.

[6] K. Choroś, Fuzzy model of multimedia retrieval system and multilevel queries. In: Information Systems Architecture and Technology. Information Technology and Web Engineering: Models, Concepts & Challenges, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, 2007, 127-134.

[7] A. Del Bimbo, Visual Information Retrieval, Morgan Kaufmann Publishers, San Francisco, 1999.

[8] N. Dimitrova, H.-J. Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor, Applications of video-content analysis and retrieval, IEEE Multimedia 9 (2002), 42–55.

[9] F. Idris and S. Panchanathan, Review of Image and Video Indexing Techniques, Journal of Visual Communication and Image Representation 8 (1997), 146–166.

[10] K. Jung, K.I. Kim, K.A. Jain, Text information extraction in images and video: a survey, Pattern Recognition 37 (2004), 977-997.

[11] I. Koprinska and S. Carrato, Temporal video segmentation: A survey, Signal Processing: Image Communication 16 (2001), 477-500.

[12] S. Lefèvre, K. Holler, and N. Vincent, A review of real-time segmentation of uncompressed video sequences for content-based search and retrieval, Real-Time Imaging 9 (2003), 73-98.

[13] M.S. Lew, N. Sebe, P.C. Gardner, Principles of visual information retrieval, Springer-Verlag, London, 2001.

[14] A.G. Money and H. Agius, Video summarisation: A conceptual framework and survey of the state of the art, Journal of Visual Communication and Image Representation 19 (2008), 121-143.

[15] S. Porter, M. Mirmehdi, and B. Thomas, Temporal video segmentation and classification of edit effects, Image and Vision Computing 21 (2003), 1097-1106.

[16] N. Sebe, M.S. Lew, and A.W.M. Smeulders, Video retrieval and summarization, Computer Vision and Image Understanding 92 (2003), 141-146.

[17] A.F. Smeaton and P. Browne, A usage study of retrieval modalities for video shot retrieval, Information Processing and Management 42 (2006), 1330-1344.

[18] A.F. Smeaton, Techniques used and open challenges to the analysis, indexing and retrieval of digital video, Information Systems 32 (2007), 545-559.

[19] D. Yamamoto, K. Nagao, iVAS: Web-based video annotation system and its applications, 2004, International Semantic Web Conference (iswc2004.semanticweb.org/demos/29/paper.pdf).

[20] H.-W. Yoo, Retrieval of movie scenes by semantic matrix and automatic feature weight update, Expert Systems with Applications 34 (2008), 2382-2395.

[21] H.J. Zhang, A. Kankanhalli, and S.W. Smoliar, Automatic partioning of video, Multimedia Systems 1 (1993), 10-28.

[22] H. Zhang, C.Y. Low, and S.W. Smoliar, Video parsing and browsing using compressed data, Multimedia Tools Applications 1 (1995), 89–111.

[23] H. Zhang, J.E. Fritts, and S.A. Goldman, Image segmentation evaluation: A survey of unsupervised methods, Computer Vision and Image Understanding 110 (2008), 260-280.

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-46

Localizing and Extracting Caption in News Video Using Multi-Frame Average

Jinlin GUO, Songyang LAO, Haitao LIU, and Jiang BU
School of Information Systems and Management
National University of Defence Technology NUDT, ChangSha, Hunan Province, China
e-mail: gjlin99@yahoo.com.cn

Abstract: News video is a very important video source. Caption in a news video can help us to understand the semantics of video content directly. A caption localization and extraction approach for news video will be proposed. This approach applies a new Multi-Frame Average (MFA) method to reduce the complexity of the background of the image. A time-based average pixel value search is employed and a Canny edge detection is performed to get the edge map. Then, a horizontal scan and a vertical scan on this edge map are used to obtain the top, bottom, left and right boundaries of the rectangles of candidate captions. Then, some rules are applied to confirm the caption. Experimental results show that the proposed approach can reduce the background complexity in most cases, and achieves a high precision and recall. Finally, we analyze the relationship between background variation of frame sequence and detection performance in detail.

Keywords. news videos, caption detection, multi-frame average, edge detection

Introduction

News video is one of the most important media and it plays a very important role in obtaining information. With hundreds of thousands of hours of news videos, there is an urgent demand for tools that will allow us efficient browsing and retrieving of video data [1]. In response to such needs, various video content analysis techniques using one or a combination of image, audio, and textual information present in video have been proposed to parse, index, and abstract massive amounts of data [1], [2]. Among these information sources, caption present in the news video frames usually annotate information on where, when, and who reported this event, and plays an important role in understanding the content of a news video sequence quickly. In summary, captions in news video frames provide highly condensed information about the content of the video and can be used for video skimming, browsing, and retrieval in large news video databases. Therefore, there will be great significance if captions in news videos can be localized, extracted, and automatically recognized.

Most of the published methods for caption extraction in videos can be classified into two categories. One category is extracting captions in individual frames independently [3], [5], [9]. The other category is utilizing the temporality of video sequences [6], [7], [8]. The first category can be further divided into two kinds: connected component-based method [9], which may have difficulties when captions are embedded in complex background or touch other objects, and texture-analysis-based

method [6], which can be very sensitive to font sizes and styles and also accurate boundaries of caption areas are hard to be found.

Existing methods do some solve some problems to a certain extent, but not perfectly. One of the key difficulties in caption localization comes from the complexity of background. Multi-Frame Integration techniques have been employed to reduced the influence of the complex background.

Hua et al. [6] use multiple frame verification to reduce text detection alarms. Then, choose those frames in which the text is most likely clear. And detect and joint every clear text block from those frames to form a clearer “man-made” frame. Hua`s method is very time-consuming and not fit for extracting characters that are made up of horizontal, vertical, and diagonal lines such as Chinese. Wang et al. [7] employ a Multi-Frame integration method, i.e. time-based minimum (or maximum) pixel value search to obtain the integrated images for the purpose of minimizing the variation of the background of the image. Wang`s method cannot reduce the complexity of background when the pixel value we want is between the biggest one and the smallest one at the corresponding pixel position.

Here, we present a novel video caption localization and extraction approach using Multi-Frame Average (MFA) after analyzing the appearance characteristic of caption in news videos. Different from previous method, we perform MFA before caption localization process, which can reduce the complexity of the whole background of the frame and make caption localization and extraction much easier.

1. Caption Localization and Extraction Algorithm

It can be found from real-life news videos always contain immobile captions and captions always appear at least 2 seconds for better understanding, i.e. at least 50 consecutive frames. Further, captions have some characteristics as follows: caption always appears in the lower area of the screen; caption is aligned horizontally; and caption usually has a good contrast from the background.

Shot is a clip that is continuously recorded without breaks, and usually is regarded as the basic unit in video information processing. So, we first segment the video into shots using the shot detection method [4]. Then, our caption localization and extraction system is performed every 25 frames within each shot. Before applying caption localization and extraction on one frame, MFA is applied on this frame and its consecutive 24 frames. The reason why we select 25 frames for average is that it guarantees that the same caption exists in consecutives 25 frames at least once.

Figure 1 shows the flow chart of our caption localization and extraction approach. Video streams are first decompressed into individual frames and frame images are taken as input. MFA is first applied and finally accurate caption bounding rectangle is extracted as to be shown in this section.

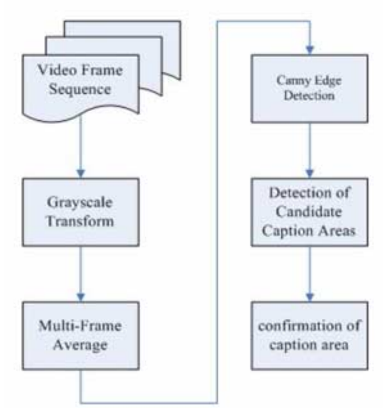


Figure 1. Flowchart of caption localization and extraction algorithm

1.1. Greyscale Transform and Multi-Frame Average

Firstly, we convert the colourful input images to grey images. For each frame in a frame cluster S_i (from i_{th} frame to $(i+24)_{th}$ frame), greyscale transform is performed on it according to the formula as follow:

$$L_i(x,y) = 0.299 \times R_i(x,y) + 0.587 \times G_i(x,y) + 0.114 \times B_i(x,y)$$

where $R_i(x,y)$, $G_i(x,y)$, $B_i(x,y)$ are the R, G, B values of the pixel of position (x, y) of i_{th} frame respectively. $L_i(x,y)$ is the greyscale.

MFA is performed on the 25 grey images obtained above. For S_i , we generate one image as follow:

$$Ave\,\mathrm{Im}\,age_i(x,y) = \frac{\sum_{j=i}^{i+24}(L_j(x,y))}{25}$$

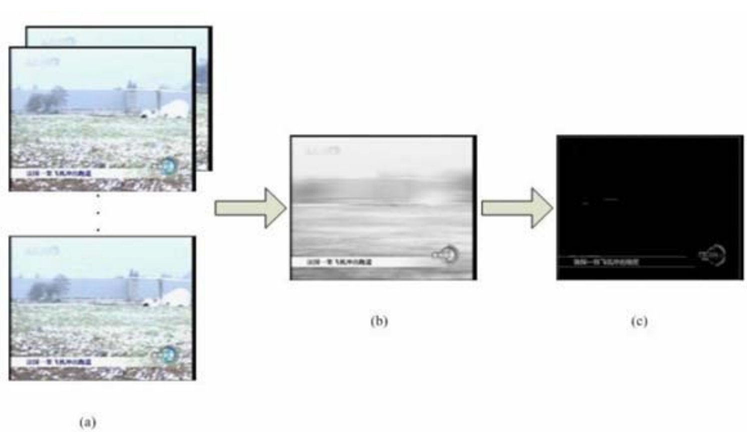


Figure 2. Example of Multi-Frame Average method. (a) Frame sequence input, (b) AveImage after Multi-Frame Average (c) Edge map of AveImage

The next step will detect candidate caption areas.

1.2. Edge Detection and Localizing Candidate Caption Areas

In section 1.1, for every frame cluster s_i, we generate one image: AveImage. Then, we will employ Canny edge detection [10] on the image.

Candidate caption area localization will be performed on the edge map obtained above. We will use a rectangle to denote the localized candidate caption area, which is similar to the method in [3]. So, this step is how to obtain the top, bottom, left, right boundaries of the rectangle.

Step 1: Obtain the width and height of the input map, and denoted as W, H respectively.

Step 2: Scan from bottom to top, from left to right on the map horizontally. Count the number of pixel whose greyscale is 255 of each line, denoted as *LineCount*. If *LineCount>T1*W*, so this line may be caption line, then, set the *IsCaptionLine* true.

Step 3: Scan from bottom to top, from left to right the map horizontally. If the number of continuous caption line is bigger T2, then, we get the top and bottom positions of these continuous caption lines as the top and bottom of the candidate caption area, denoted as *top* and *bottom*, go to Step 4. Otherwise, go on to scan the map from next line of *top* horizontally, repeat Step 3 until the scan is over.

Step 4: Scan from left to right, from bottom to top in the area between the bottom and top vertically, that is, count the number of pixel whose greyscale is 255, denoted as *ColumCount*. If *ColumCount>T3*(top-bottom)*, so the line is considered as caption column, and set the *IsCaptionColum* true.

Step 5: From left to right, from bottom to top, scan the area between the bottom and the top vertically. If the number of continuous caption columns is bigger than T4, than we get the left and right positions of these caption columns, denoted as *left* and *right* respectively. Back to Step 3, and go on to scan the map from next line of *top* horizontally.

Where *T1*, *T2*, *T3*, and *T4* are threshold, and we set *T1=0.1*, *T2=5*, *T3=0.2*, and *T4=10* experimentally.

1.3. Confirmation and Extraction of Caption Area

Then, some rules are applied to remove some false alarms according to the appearance rule of captions:

(1) Caption in news video always appears in the lower area of the screen, so *top* is smaller than 1/4*H;

(2) The horizontal-vertical aspect ratio of the caption rectangle must be bigger than 2.0;

(3) The distance between the bottom (or top, left, right) boundary of caption rectangle and the image boundary must be bigger than 10 pixels;

(4) Height of caption rectangle must be bigger than 9 pixels.

The caption rectangle that does not satisfy one or more of above rules will be removed.

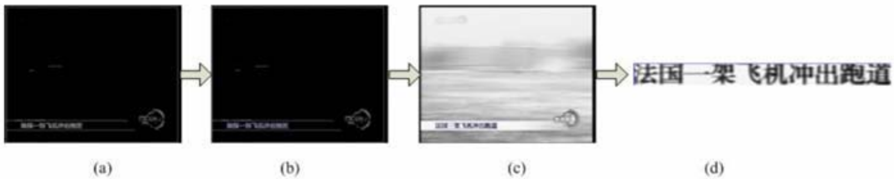


Figure 3. Choosing caption location and extraction on the Canny edge map of AveImage. (a) edge map of AveImage, (b) candidate caption area, (c) caption area location on the AveImage after filtering, (d) caption after extracting

2. Experiments

We have tested our method on a test set with 100 news clips from CCTV1 of a resolution of 720×576, and each clip consists of several consecutive frames containing the same steady captions.

Recall and precision are used to evaluate the performance of the whole algorithm. They are defined as:

$$Recall = \frac{a}{c}$$

$$Precision = \frac{a}{a+b}$$

Where a is the number of captions detected as captions, b is the number of non-captions detected as captions, and c is the total number of truly captions in the test set.

The caption localization and extraction approach is applied on each clip and MFA is performed. The approach without MFA is also applied on the first frame of each clip. Performance comparison among our algorithm, the approach and the method in [5] are listed in Table 1.

Table 1. Comparison of results of caption localization extraction algorithms

Method	Total caption rectangles	Detected	False alarms	Recall	Precision
MFA	100	105	5	100%	95.2%
Non-MFA	100	126	26	100%	79.4%
Method of [5]	100	117	16	100%	85.5%

From Table 1, we can see that our MFA technique can significantly reduce the false alarms. The reason is that the improvement of the quality of background can enhance the contrast of some low-contrast caption rectangles and makes caption localization much easier.

It can be seen that the caption rectangles are quite tight and accurate from Figure 3.

Sometimes, the MFA method can't improve the background complexity. So, we perform another experiment on 40 video clips chosen from the test set above randomly to analyze the relationship between background variation of frame sequence and detection performance.

For the i_{th} frame and $(i+1)_{th}$ frame of the 25 greyscale images input:

$$N_{i,x,y} = \begin{cases} 1 & \text{if} \quad |f(x,y,i+1) - f(x,y,i)| > T \\ 0 & \text{else} \end{cases}$$

$$N = \sum_{i=1}^{24} \sum_{y=1}^{H} \sum_{x=1}^{W} N_{i,x,y}$$

where $f(x,y,i)$ is the greyscale of (x,y) of i_{th} frame; T is threshold and set as 50; N represents the variation of background of frame sequence.

N_{MFA} / N_{TA} is used to evaluate the performance of reducing the complexity of background, where N_{MFA} is the number of edge points in the average frame and N_{TA} is the average of number of edge points of 25 images.

Experimental result is given in Figure 4. The curve is 3-order polynomial fit.

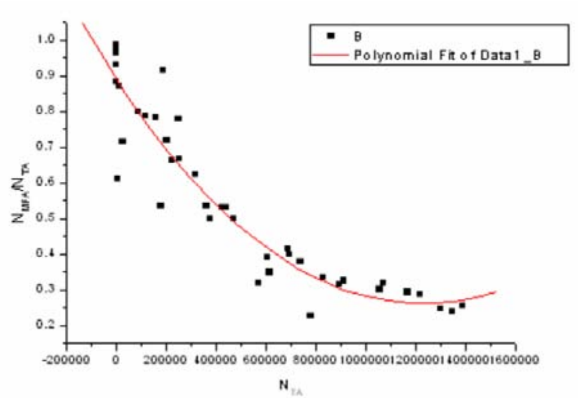


Figure 4. Experimental result of relationship between background variation of frame sequence and detection performance

From Figure 4 we can get such conclusions:

(1) MFA method can improve the quality of background and is helpful in localization of captions;

(2) when variation of background of frame sequence is big enough, MFA can reduce complexity of background highly.

3. Conclusions

A news video caption localization and extraction approach using Multi-Frame Average has been presented.

There are three highlights in our proposed method. First, we proposed a novel MFA technique for caption localization and extraction, which is applied on video frames before caption localization process. Experimental results show that it can significantly reduce false alarm and improve the precision of detection. Second, we used a horizontal-vertical scan method, which can obtain tighter and more accurate caption rectangle of caption areas. Third, we analyzed the relationship between background variation of frame sequence and detection performance in detail.

Acknowledge

The research described has been supported by NSF (60572137) and National 863 Plans Fund (2006AA01Z316) of China.

References

[1] M. Christel, S. Steven, and H. Wactlar, Informedia digital video library, *Proc. ACM Multimedia Conf.,* Oct. 1994, 480-481.

[2] A. Hauptmann and M. Smith, Text, speech, and vision for video segmentation: The Informedia Project, *AAAI Symp. Computational Models for Integrating Language and Vision*, 1995.

[3] XIE Yu-xiang, LUAN Xi-dao, WU Lin-da, and LAO Song-yang., Caption detection in news video frames. *J. Computer Engineering* 30 (2004), 167-168.

[4] ZHU Xiao-jun, LAO Song-yang, WU Zuo-shun, and XIE Yu-xiang, An effective shot detection method, *[J]. Computer Engineering and Application* 39 (2003), 59-61.

[5] R. Lienhart and F. Stuber, Automatic text recognition in digital videos, *Proc. of SPIE*[C].[S1]:[sn],1996, 180-188.

[6] X. Hua, P. Yin, and H.J. Zhang, Efficient video text recognition using multiple frames integration, *IEEE Int't Conf. Image Processing (ICIP)* , Rochester, New York, 2002.

[7] RongRong Wang, JunJun Wan, and LiDe Wu, A novel video caption detection approach using Multi-Frame Integration, *IEEE Proceeding of the 17th International Conference on Pattern Recognition (ICPR*), 2004.

[8] Xi J., et al., A video text detection and recognition system, *Proc. of ICME 2001*, 1080-1083, Waseda University, Tokyo, Japan, August, 2001.

[9] A.K. Jain and B. Yu, Automatic text location in images and video frames, *Pattern Recognition* 31 (1998), 2055-2076.

[10] YAO Ming, et al., *Digital Image Processing*, BeiJing, China Machine Press, 2006.

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-53

SMiLE - Session Mobility in mobiLe Environments

Günther HÖLBLING [a] Wolfgang PFNÜR [a] Harald KOSCH [a]

[a] *{hoelblin,pfnuer,kosch}@fim.uni-passau.de,*
Chair of Distributed Information Systems, University of Passau, Innstraße 43, Passau, Germany

Abstract. Nowadays most people own several devices, like a notebook or a smartphone, for satisfying their mobility and flexibility needs. Without special arrangements a program, its execution state and saved files - the latter two are commonly called "session" - are confined to a physical host. Session Mobility enables the user to break this law of locality, and instead creates a relation between the session and himself. In this paper we present a novel system for supporting session mobility in various scenarios. To support an almost automatic session handover between different devices a mobile agent system has been used. The selection of target devices has also been automated based on the usage context, the device's capabilities and a rough estimation of the actual location. Based on the Session Mobility in mobiLe Environments platform (SMiLE) a use case has been realized where a video session is migrated from a notebook to a smartphone and vice versa.

Keywords. session mobility, session handover, Agent System, JADE, MPEG-21

Introduction

Today most users are in the unpleasant situation of having their programs and data spread over several devices. To keep the appropriate data on the appropriate device available becomes a nuisance. To be able to continue the actual work on another system without losing much time, an automated system is needed. Session Mobility enables the user to detach a session from an actual device. The user can access his data on any physical host that is part of the SMiLE platform. The system utilizes the technology of mobile agents to meet the user's need for mobility. Thus the user does not have to care about the location of the actual session of his work, instead the session takes care of the user's location. For compatibility reasons and assuring future prospects a standardized format - MPEG-21 - for representing the actual session has been used.

usage examples Considering a scenario where someone is sitting in the living room in front of his TV-set and watching football: As nothing important seems to happen, that person decides to go to watch the game with a good friend only two blocks away. After leaving his house he hears a multitude of screams from his neighbors. When he arrives at his friend he has to realize that he missed the decisive goal. By the use of SMiLE that person could have continued watching or at least hearing the broadcast on his PDA or smartphone by moving the actual session from his settop-box to his mobile device. By

tracking the location of the user the session migration to the mobile device could be done automatically when the user is on the move.
Another usage example of the system would be mobility of "browsing sessions". In the simplest form a "browsing session" could be made up by a collection of websites and their associated metadata (e.g. their access dates or cookies). It enables for example a user to move the active session from his PC to his smartphone. Thus the user is able to continue scouring the web on the way.
The rest of the paper is organized as follows. The sections 1 and 2 give background information on the MPEG-21 standard and on mobile agent systems. Section 4 presents our system in detail. Besides the architecture, the selection of the target hosts based on a benchmark and the session migration will be discussed as well. In section 5, we take a look at several similar projects and discuss the main differences compared to our proposal. Finally, section 6 concludes the paper with a short summary and future work.

1. MPEG-21

MPEG-21 tries to realize a standard for the "big picture" of multimedia systems and specifies a framework to create a common base for "seamless and universal delivery of multimedia".[5] MPEG-21 is centered around 7 key areas:

1. Digital Item Declaration
2. Digital Item Identification and Description
3. Content Handling and Usage
4. Intellectual Property Management and Protection
5. Terminals and Networks
6. Content Representation
7. Event Reporting

For the session declaration and mobility focused in SMiLE only the areas Digital Item Declaration and Terminals and Networks are of further relevance in this context. Nevertheless other applications providing presentation or playback functions have to cover also most other MPEG-21 key areas.

1.1. Content DI and Context DI

SMiLE makes use of two MPEG-21 specifications for describing the actual content and its associated session, the Content Digital Item (Content DI) and the Context Digital Item (Context DI)[8]. Both of them are XML files that use MPEG-21 schema definitions. The Content DI represents the actual content in form of ressources, the metadata and their interrelationships. It may also offer several choices and selections for a content each providing a different quality, format, mediatype or resolution. Thus a Content DI can be consumed on different devices in an adapted way. For example a Content DI may contain a movie at 3 chooseable qualities and transcript of the movie for devices without video playback capabilities.

The Context DI saves the actual session information. It contains information about the playback conditions, e.g. what choices were made to play back the Content DI or the actual playbacktime. The session is saved within a Digital Item Adaption Description

Unit of type SessionMobilityAppInfo. This description type is used to adapt the Content Digital Item, it also stores information about the application that uses that content. As the information that is needed to preserve a session vary from application to application, the format of the session description is not standardized. [4]

1.2. Terminal Capabilities

For describing the capabilities of different devices SMiLE adopts the mechanism of MPEG-21 for characterizing terminals. All relevant properties of the hardware are saved in an MPEG-21 conformant format. These capabilities include Benchmark (for CPU), PowerCharacteristics (for battery time), Storages (for RAM size), DeviceClass (PC, laptop, PDA, ...), Displays (resolution and bit-depth), AudioOutputs (channels, bits per sample, sampling frequency) and CodecCapabilities for Audio and Video Codecs.

2. Agent Systems

A Software Agent is in general a program that carries out tasks and makes decisions on behalf of a user or an other program. An example would be a searchbot, that scours the web for useful information. It decides wether or not a website has relevant information, and extracts that knowledge for the user. Agents are often defined by their characteristics - namely autonomy, proactivity, reactivity, adaptivity, persistence and social-ability. Autonomy means that the agent can act autonomously without need for user interaction and proactivity means that it can act out of his own will. The agent still may have to ask for permission before doing anything that might be potentially harmful. The agent can react to its environment and adapts to changes in it - this is called reactivity. Therefore it may need to gather information about its surrounding. Persistence means that agents continuously run and are not stopped after their task has been finished. The ability of agents to communicate and cooperate with other agents and components is called social ability. Besides these characteristics mobile agents have the ability to move from the current to another host. This action is called “migration”. For a detailed explanation of agent systems see [10,1].

Several agent systems have been evaluated as basis for SMiLE. Because of the need of mobility and the support of mobile devices e.g. handhelds and smartphones we focused on Java based platforms. These include Aglets, Beegent, Bond, JACK, MIA, UbiMAS, Mole, Voyager, Grasshoper, Gypsy, Cougaar, Agent Factory and JADE. Most of these agent systems were discarded for being out-dated or the lack of supporting mobile devices (especially smartphones). The most promising platforms were JADE and Agent Factory. Finally JADE was chosen for its mature status and the broad support for different Java runtimes like J2ME in both configurations (CDC and CLDC), Personal Java and J2SE. For comparative discussion of several mobile agent platforms, the reader can refer to [15].

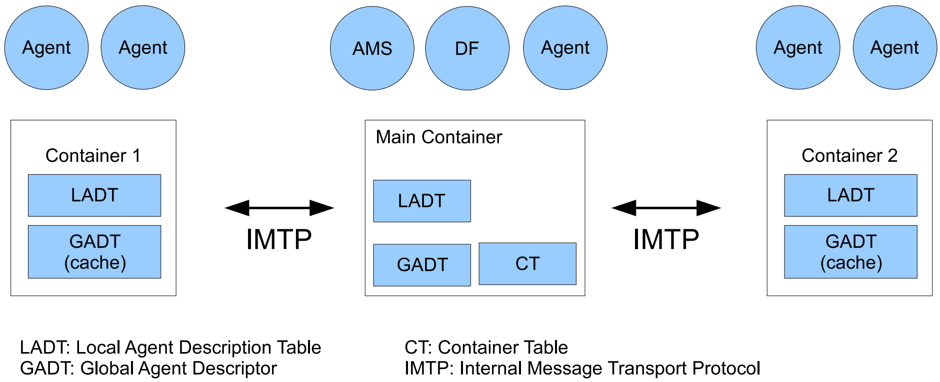


Figure 1. JADE organizational structure.

3. JADE

JADE (Java Agent Development Framework)[3] was developed and distributed by Telecom Italia, but published under the LGPL2 (Lesser General Public License Version 2). It is a Java based agent framework which fully complies to FIPA[1] standards[7].

A JADE platform is an execution environment for agents that may spread over several physical hosts. It consists of several agent containers, that can run on different machines. One of these containers is the main container, which hosts the organizational information and processes. All standard containers have to register at a main container, and there is only one active main container per Agent Platform. The organizational information and processes hosted by the main container are as follows (cf. figure 1):

- The container table (CT), which is the registry where the transport addresses and object references of all containers of the platform are saved
- The Global Agent Descriptor Table (GADT), where all agents of the platform register themselves, including status and location
- Hosting of the Agent Management System agent (AMS) which takes care of most organizational tasks like registering new agents, deregistering agents upon deletion and taking care of the whole migration process.
- Hosting of the Directory Facilitator agent (DF) which provides a registry for services provided by agents. This mechanism is similar to the yellow pages of UDDI.

The main container is the central component of the platform. It hosts the most important agents, the Directory Facilitator (DF) and the Agent Management System (AMS) agent which are only present on the main container. Nevertheless most operations do not include the main container at all, as every container keeps its own copy of the GADT, which is named Local Agent Descriptor Table (LADT). If the local table is out of synch, the container will refresh its cache. Typically this happens when a queried entry does not exist, or if the querying agent reports that the information was inaccurate. Due to that agent cache the main container is only involved when agents or services are created, deleted or changed (which includes agent migration). The main container is however still

[1]The Foundation for Intelligent Physical Agents provides a collection of standards to ensure interoperability between different FIPA compliant agent systems.

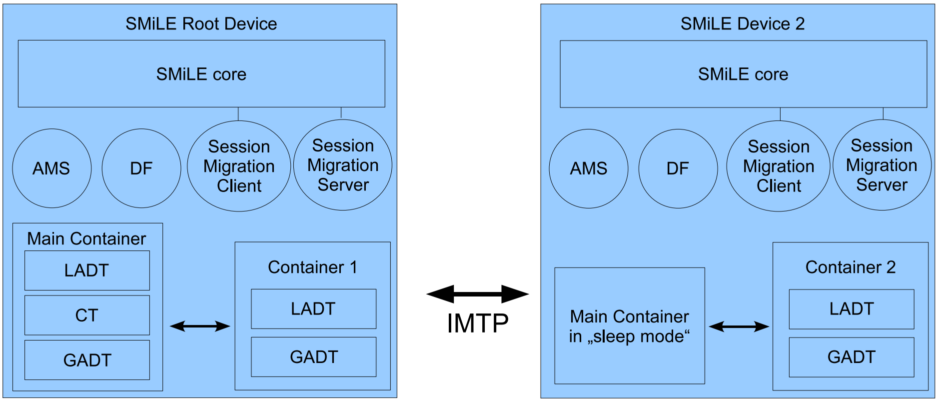


Figure 2. Overview of SMiLE's architecture.

a single point of failure, a crash would disable the agent platform. The Main Container Replication System (MCRS) helps to overcome this situation by starting several "sleeping" main containers which simply act as a proxy to the active main container. In case of a failure, the remaining main containers are notified, and reorganize accordingly.

Message Transport is implemented in two different ways: For inter-platform communication there is a http interface available at the main container. This is the FIPA-conform entry point for messages that are sent from outside into the platform.

For intra-platform communication a special protocol called Internal Message Transport Protocol (IMTP) is used, which is not FIPA-conform. Besides standard message exchange between agents it is also used for system-message exchange, for example commands to shutdown a container or kill an agent.

JADE-LEAP The limitations of the hardware of mobile devices (J2ME CLDC) made it necessary to introduce a special Lightweight Extensible Agent Platform (LEAP) for JADE. LEAP uses the so-called split-container mode: There is a lightweight front-end container on the mobile device, which keeps a constant connection to a back-end container on a J2SE host. Neither the front-end nor the back-end are a container on their own, as both only provide part of the functionality - hence the name "split-container". As serializing and deserializing of objects is not supported on J2ME CLDC, "real" migration is not possible.

Unfortunately, this also limits the ability of Agents used in SMiLE. SMiLE agents can only achieve mobility by one of two distinct means: Either they are "mobile agents" as in "running on mobile devices" (using Jade-Leap and no migration), or they are "mobile agents" as per definition, able to migrate, but unable to run on mobile devices.

4. Description of the System

SMiLE has been implemented in Java mainly to achieve platform independence and a broad support of mobile devices. Figure 2 shows a rough overview of a SMiLE device and the relation to the services provided by the agent platform. JADE is used in the system for message delivery and migration. Every device in the SMiLE platform can

offer itself as a possible migration host through JADE's DF-service (cf. section 3) by registering the migration service with its MigrationServer agent.

To enhance the fault tolerance, every J2SE instance of SMiLE has its own agent container as well as its own main container. Only one of the networked devices' main containers is active, the others are working as proxies. Based on the MCRS (cf. section 3) the main containers of different devices are organized in a ring, where every local main container keeps up a connection to the next main container. If the connection to the active main container is lost, the ring will re-organize and one of the other main containers changes into active state. This mechanism guarantees a maximum degree of fault tolerance. Without it, the whole System would disintegrate instantly if the JADE MainContainer crashes, due to the loss of most organizational information of the agent system.

On startup every new device has to register its MigrationServer agent at the Directory Facilitator of the agent platform. For that reason the first SMiLE device creates a main container and its associated "service"-agents (DF and AMS). This device now runs the first active main container. To avoid having multiple singleton SMiLE networks, every device sends a multicast request on startup. If the request is received by another instance of SMiLE, it will reply, and thus give the recently started device the chance to connect to the existing network and to its MCRS. In a final step the MigrationServer Agent is started and registered. This agent manages all incoming migration requests of other devices and negotiates the session handover. The MigrationClient agent is only started if an actual session should be moved to another device. A detailed description of the migration process can be found in section 4.2. The communication between SMiLE Devices is handled by JADE based on the Internal Message Transport Protocol (IMTP).

As depicted in figure 3 SMiLE consists of several parts that will be described in the following.

- SMiLE core: This part provides the main functionality of the system. It cares about evaluating system capabilities (see section 4.1), migration target selection (see section 4.1), the agent code (MigrationClient agent and MigrationServer agent) and the session migration process itself (see section 4.2).
- MPEG-21 Layer: All the information about content (where is the content), context (what has been done with that content) and devices (e.g. benchmark results) are saved in MPEG-21 conform XML. Thus the MPEG-21 Layer was introduced to support the creation and processing of MPEG-21 compliant XML- documents. Moreover it adopts the Least Recently Used (LRU) strategy for buffering frequently accessed DI, making access to those items more efficient.
- Virtualisation Layer: J2SE and J2ME share many classes. However there are major differences in the way user-interfaces are realized, files are accessed or which libraries are supported. Thus several functions of SMiLE, for example determing terminal capabilities (e.g. supported codecs or resolution and color-depth of the screen), or storing a digital item, have to be implemented in two editions - one for J2ME and one for J2SE. For these reasons a virtualisation layer has been created. It allows to keep as many of the classes as possible working on both editions. Basically, the classes can be compiled for both editions, while still being able to use edition-specific functions through the virtualisation layer. That way nearly all the classes except the user interface and some of the virtualisation layer's classes

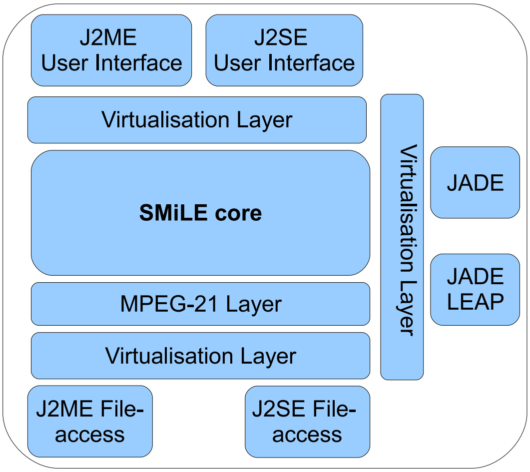


Figure 3. Overview of SMiLe's architecture.

can be shared. The access to the agent platform is managed by this layer as well, to keep version specifics out of SMiLE's core .

- J2ME/J2SE specific parts: There are several edition specific parts in SMiLE like the user interfaces, file access or the agent platform (see section 3)

4.1. Benchmarking and Hardware Evaluation

To be able to decide whether or not a device is fast enough to continue a specific session, the hardware capabilities of the machine and a scale for comparison was created. Three main categories were measured: CPU speed, resolution and color capabilities.

Additionally a session specific category, the multimedia capabilities of the devices, was evaluated. Thus for most devices a list of supported audio and video CoDecs is available.

CPU speed: The CPU speed was measured by calculating a total of 1 million additions, multiplications and divisions. This will take a couple of seconds on a reasonably fast mobile phone (e.g. Nokia N80), and is very fast (split-second) on a personal computer. The time taken for this test is measured in milliseconds, with a maximum of 20 seconds for the test.

For slow devices, a good resolution can be achieved by simply using a linear function: For every 0.2 seconds that were needed to complete the test, 1 point is subtracted from a starting value of 100 points.

Unfortunately, this means that fast devices that need less than 1 second for the test (which should be most personal computers) will all receive nearly the same rating. For fast devices a function based on indirect proportionality between points and time taken would result in a high resolution, but for slow devices, the resolution would be very bad since e.g. devices needing 10.2 to 20 seconds would all receive 1 point.

By taking the average of both functions (f1 and f2), an acceptable resolution was achieved for both fast and slow devices (f3).

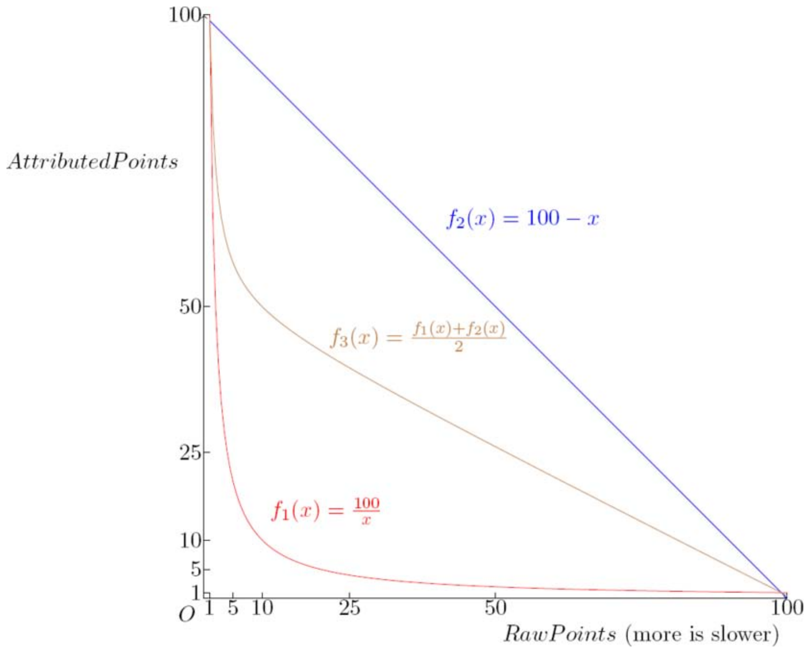


Figure 4. CPU raw points - attributed points diagram

The three functions can be seen at figure 4. A logarithmic scale might have been able to perform similar, but J2ME CDC 1.0 does not support logarithmic or exponential functions.

Resolution and the number of colors: As displays are 2-dimensional, the squareroot of the number of pixels was used as a basis for attributing points. A normalizing factor was introduced to grant 100 points to all displays exceeding 2 MPixel (2000*1000), displays with 200 pixels (20*10) would still achieve one point.

As the number of colors is usually dependant on the number of bits used to identify a specific color, this value is used for calculating the points. This allows to compare 8-bit colors with 32-bit colors. As 32-bit is the current standard for desktops, a normalizing factor was introduced so a colordepth of 32-bit results in 100 points.

To prevent black-and-white displays with a high number of grayscale colors to get a better result than a simple color display, a punishing factor of 5 was used. If a display is only black-and-white, it gets only 1/5 of the points that a color display with an equal number of colors would get.

Priorities and Decision: Thanks to the pre-processing described above, there are 3 point-values each ranging from 1 to 100 points that describe the underlying hardware. Those 3 values are not enough to make a simple decision: Unless all 3 values of one benchmark are higher than those of the other, weighting factors are needed. These weighting factors should depend on the application used e.g. for pictures, colors are important, for large data-sets like tables screensize and thus resolution is important, and for some tasks like decryption CPU power is most important. For videos, all of these are important - but only up to a certain point. If the Codec is supported and the CPU is fast enough to decode the video in real-time, more

CPU power will not yield further improvements. Still, if the computing power is too low, the video will stutter.

What is needed is a single value that represents the ability of the underlying hardware to perform a specific task, like playing a video, starting a game, or similar. This is done by designing special priority classes for that action, which can subsequently be used for calculating a prioritized points value.

Priorities basically assign an importance-value to every of the three main categories that were measured. The idea behind those priority-values is that they shall specify the minimum number of points needed to deliver the best service possible. If there is still an improvement from 49 to 50, but not from 50 to 51, then the priority should be 50.

To calculate a single value out of the 3 categories' values, each of them is divided by the priority value that was assigned to that catgory. That way, a fulfillment ratio for every of the three main aspects is created. The final value is then calculated using the bottleneck principle, as it does not matter how fast the CPU is and how brilliant the colors are if the screen is too small to recognize anything. Thus, the smallest of these three aptitude ratios determines the final points value: it is simply multiplied by 100. If a category was assigned an importance of 0 (no importance), this category is simply ignored. If all three categories are assigned an importance of 0, every system will get 100 points.

The result of these calculations is then used for automatic decisions. To be able to do this, every Content Digital Item is assigned one point value (and a corresponding priority) per component. This value is then compared with the points that were calculated for that priority during benchmark.

4.2. Session Migration

As mentioned before, the session information itself is saved and transferred as Digital Items in MPEG-21 format. There are several agents involved in the migration process. Figure 5 shows the participants and the migration process in a simplified form. As migration is not fully supported on J2ME since it does not support serialization and reflection, there are two editions for client- and server applications. Depending on wether the actual device is the initiator or the target of a migration the corresponding application has to be started - MigrationClient for migrating to another place, or MigrationServer to accept a migration. Both components create a corresponding agent on startup. To provide a seamless migration the MigrationServer is typically started in the background, waiting for migration requests. When both the client and the server are running, the migration process can be initiated. If no benchmark and capability information is present on startup of the MigrationServer a benchmark is forced. The default means of communication within SMiLE is the JADE IMTP system, which is used by all agents. The connection between different devices is established using WLAN. A typical migration consists of 7 steps:

Step 1 To become available on the platform for migration requests the MigrationServer has to be registered at the DF in the first place.

Step 2 For discovering all available migration targets the MigrationClient agent queries the Migration Helper agent.

Step 3 and 4 The Migration Helper Agent first queries the Directory Facilitator (DF) to get a list of all available hosts. Then a bluetooth scan, saving all bluetooth identifiers found, is performed. By comparing the bluetooth IDs of the available hosts with the list of bluetooth IDs the hosts are marked as reachable or unreachable. The bluetooth interface is only used to gain a rough estimate on the distance to a possible target.

Step 5 In this step the list of hosts with the additional reachability information is returned to the MigrationClient agent.

Step 6 The host selection can be done automatically or manually. In both variants the benchmark results and the capabilities of each available host are taken in account to identify the best host for the actual session. Different session types also lead to different decisions by providing different priority factors (see 4.1). For example automatic migration of a video session will simply pick the host with the highest number of Video-Points and initiates migration. In manual mode it is possible to choose from a complete list of all hosts and the number of points they achieved. It is also visible wether or not this client is within bluetooth range, so decisions can be based on locality. To be able to distinguish the hosts, their hostname is used for J2SE hosts, and their telephone number for J2ME hosts. As the telephone number is not available to a MIDlet for security reasons, it has to be manually specified in the config.xml. In addition to the actual session other digital items of the client device can be marked for migration as well.

Step 7 Due to restrictions of JADE-LEAP "real" migration is only supported between J2SE hosts. Thus our solution to provide session migration on J2ME is to send working instructions specifying the task and the session data to a specialized agent on the server side. This agent has to provide the implementation for these instructions, otherwise the migration fails. After the migration is initiated, the digital items are sent to the new host. When all digital items are transmitted, a message is sent to indicate that the migration is complete. During a migration, all migration related messages from different hosts are discarded.

As soon as a migration is finished the MigrationServer continues the specified session.

4.3. Test Setup

The system was tested with a video-session scenario where two smartphone (a Nokia N80 and a Nokia N95) and a standard laptop were used. All devices were equipped with a bluetooth interface and connected to a Wireless Network. For video streaming the Darwin Streaming Server was used. A video was transcoded as .3gp file for the cellular in a low quality and in a higher quality for standard PC's. The 3gp version of the video was started on the cellular (Nokia N95), paused and the session migration process initiated. According to the steps described in 4.2 the session information was saved as an MPEG-21 Digital Item. By evaluating the benchmark results of the available devices the laptop is chosen as migration target due to the superior benchmark results compared to those of the other smartphone. After the session was migrated to the laptop, the higher quality version of the video is continued at the very spot it was stopped on the cellular. While both versions of the video were referenced within the Content Digital Item, the benchmark results of the laptop resulted in the selection of the higher quality version, while the smartphone was only capable for the playback of the low quality version.

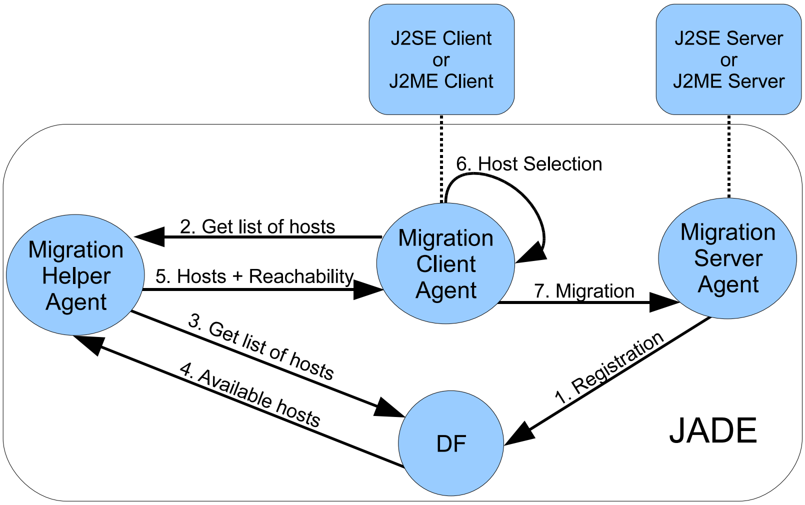


Figure 5. Migration Process.

5. Related Work

Several approaches have been made to satisfy the users needs for mobility and flexibility in their daily work. A recent and very simple form of session mobility can be obtained by using portable software on USB-sticks (e.g. The PortableApps Suite). All programs that should be "mobile" available have to be installed directly on the USB-stick. Although this approach is quite practical on a PC-platform it can not be used with most mobile devices, especially not with mobile phones. In [9] a streaming system for ubiquitous environments has been defined. Based on MPEG-21 it provides a mechanism for session mobility. Thus users are enabled to continuously consume media seamless through several terminals. Moreover the system is able to transcode media or adapt it to the user's environment. Even though this proposal seems to be very similar to SMiLE several major differences can be identified. Compared to our proposal the system focuses on streaming media and is only capable of transferring "media" sessions. It also uses a central server for parsing metadata, managing DI's and user sessions. The Internet Indirection Infrastructure (i3) is used in [16] to achieve Personal Mobility. It suggests to use bluetooth-identifiers to locate persons. Every device that is within reach of the bluetooth-identifier of a specific person will try to register by setting an i3 trigger that represents that user. If multiple devices are within reach, an internal protocol will decide which one will handle the actual session. Compared to SMiLE, there is no application-specific algorithm that will decide which device to use, and the user cannot choose manually. It is not possible to use applications that are not using a client-server model, as there is no direct transfer of the session. Several Session Initiation Protocol (SIP) based systems for mobility for multimedia applications have been proposed. In [12] the main scenarios are re-routing of calls and streaming video depending on the location of the user. Sajal et. al. [2] proposed an architecture for mobility management supporting soft hand-off for voice or video streams. While these proposals certainly share some functionality with SMiLE they are focused on streaming media and there is no task-based algorithm that helps decide which device to use. [13] uses agents to transfer sessions. Sessions are

saved in XML format, and task-based agents allow to perform specific tasks. The agents also take care of their own sessions, saving and loading them as needed. Compared to SMiLE, there is always a special agent per task, and there is no algorithm to help decide which device to use. Cyeon et. al. present in [6] a distributed multimedia communication framework and conferencing software where session information is managed within a LDAP directory server. This work uses a central server for session management and focuses on multimedia content and video streaming in an e-learning environment. The browser session preservation and migration (BSPM) infrastructure [14] enables the user to switch devices while browsing the web and continue the same active web session on the new device. It uses a proxy server to store a snapshot of the browser session and to provide it to other devices. In comparison to SMiLE, BSPM is only applicable for browser sessions. Moreover the system does not take the capabilities of the target device into account.

6. Conclusion

As more and more different devices are used by one person keeping the appropriate data on the appropriate device available becomes a nuisance. In this paper we presented a novel architecture for supporting session mobility. By the use of the agent platform JADE actual user sessions may be easily carried within the agent from one system to another. Based on MPEG-21 a standardized way has been found to describe the session data and its context. Besides the architecture, usable for different session types, a strategy for the selection of the "best" device for continuing a video session and a mechanism for migrating the actual session to the target device has been realized.

Our future work in this domain will concentrate on the extension of the device profiles based on the free open source project WURFL [11]. WURFL provides detailed device profiles of more than 7000 different devices. By providing better device profiles the selection strategy for the target device will be enhanced. In future we are planing to integrate support for other session types (e.g. "browser-sessions" - the bookmarks and the open tabs of the browser) in SMiLE.

References

[1] F. Bagci. Reflektive mobile Agenten in ubiquitären Systemen. `http://www.opus-bayern.de/uni-augsburg/volltexte/2006/209/pdf/diss.pdf`, 2006.

[2] N. Banerjee, A. Acharya, and S. Das. Seamless SIP-based mobility for multimedia applications. *Network, IEEE*, 20(2):6–13, March-April 2006.

[3] F. Bellifemine, G. Caire, and D. Greenwood. *Developing multi-agent systems with JADE.* John Wiley & Sons, 2007.

[4] I. Burnett, S. Davis, and G. Drury. Mpeg-21 digital item declaration and identification-principles and compression. *Multimedia, IEEE Transactions on*, 7(3):400–407, 2005.

[5] I. S. Burnett, F. Pereira, R. V. de Walle, and R. Koenen. *The MPEG-21 Book.* John Wiley & Sons, 2006.

[6] H. Cyeon, T. Schmidt, M. Wahlisch, M. Palkow, and H. Regensburg. A distributed multimedia communication system and its applications to e-learning. *Consumer Electronics, 2004 IEEE International Symposium on*, pages 425–429, 2004.

[7] F. for Intelligent Physical Agents. Standard status specifications. `http://www.fipa.org/repository/standardspecs.html`. [online, accessed 12-March-2008].

[8] ISO/IED. Information technology – multimedia framework (mpeg-21) – part 2: Digital item declaration. 2005.

[9] O. Min, J. Kim, and M. Kim. Design of an adaptive streaming system in ubiquitous environment. *Advanced Communication Technology, 2006. ICACT 2006. The 8th International Conference*, 2:4 pp.–, 2006.

[10] V. L. Padgham and M. Winikoff. *Developing Intelligent Agent Systems: A Practical Guide*. John Wiley & Sons, 2005.

[11] L. Passani. Wurfl - the wireless universal resource file. `http://wurfl.sourceforge.net/`. [online, accessed 12-March-2008].

[12] H. Schulzrinne and E. Wedlund. Application-layer mobility using SIP. *SIGMOBILE Mob. Comput. Commun. Rev.*, 4(3):47–57, 2000.

[13] M. M. Shiaa and L. E. Liljeback. User and session mobility in a plug-and-play architecture. *IFIP WG6.7 Workshop and EUNICE Summer School*, 2002.

[14] H. Song, H. Chu, and S. Kurakake. Browser session preservation and migration, 2002.

[15] R. Trillo, S. Ilarri, and E. Mena. Comparison and performance evaluation of mobile agent platforms. In *ICAS '07: Proceedings of the Third International Conference on Autonomic and Autonomous Systems*, page 41, Washington, DC, USA, 2007. IEEE Computer Society.

[16] S. Zhuang, K. Lai, I. Stoica, R. Katz, and S. Shenker. Host mobility using an internet indirection infrastructure. *Wirel. Netw.*, 11(6):741–756, 2005.

This page intentionally left blank

Data Processing in Information Systems

This page intentionally left blank

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-69

Data Mining Approach to Analysis of Computer Logs Using New Patterns

Krzysztof CABAJ
Institute of Computer Science, Warsaw University of Technology
Nowowiejska 15/19, 00-665 Warsaw, Poland
e-mail: kcabaj@elka.pw.edu.pl

Abstract. Due to the huge amount of currently collected data, only computer methods are able to analyze it. Data Mining techniques could be used for this purpose, but most of currently used techniques discovering global patterns loose information about local changes. In this paper the new patterns are proposed: frequent events and groups of events in data stream. They have two advantages: information about local changes in distribution of patterns is obtained and the number of discovered patterns is smaller than in other methods. Described experiments prove that patterns give valuable knowledge, for example, in analysis of computer logs. Analysis of firewall logs reveals interest of user, its favourite web pages and used portals. By using described methods for analysis of HoneyPot logs, detailed knowledge about malicious code and time of its activity could be received. Additionally, information about infected machines IP addresses and authentication data is automatically discovered.

Keywords. data mining, computer networks, computer security

Introduction

The growth of storage capacity in modern computers allows collection of more and more data. Therefore, analysis of such amount of data without computer techniques is almost impossible. Methods described as Knowledge Discovery in Databases could be used for this purpose, especially, the main step of it called Data Mining. There are several techniques that could be used, for example, clustering, classification, or pattern discovery. In this paper only pattern discovery is discussed. A pattern is simple, easy understandable [4, p. 35] information which carries some previously unknown knowledge which could be used afterwards [8, p.10]. There are many types of patterns, which depend on the type of stored data. One group of patterns is connected with time stamped data, for example, transactions in market, bank account operations, or computer logs. In this kind of data some previously proposed patterns could be discovered, like sequences [2], episodes [5] or episodes rules [6]. The main disadvantage of these methods is that patterns are discovered globally in all analyzed data; so, a local change in appearance is not considered, although this information is interesting. Many shop managers may be interested in what products are preferred in which period of week or year. Using this information, some decision about an ordering or a placement of products could be made. For the investigators, who analyze electronic evidence such as telecommunication billings or network logs, knowledge

about the time period when suspected communicate frequently with some people or visited some web pages could be useful.

Presented in this paper new patterns frequent events and groups of events in data stream is a proposal of techniques which allow discovering local changes in data. Frequent event in data stream could be treated as generalization of well known, in Data Mining pattern, a frequent event [1]. Generalization adds to pattern information about time period, when an event could be considered as frequent. Analysis of stored emails, using proposed method, could discover knowledge about a time period when contacts with some people are frequent. Some events appear in data simultaneously. Information about such group of events could be valuable as well. This kind of knowledge is automatically discovered by the patters called the group of events in data stream. Reduction in number of discovered patterns is the considerable advantage of this method, which allows analysts focus only on interesting once. As a result, the process of discovery and verification of new, useful knowledge is speed up.

1. Related Work

The author of this paper introduces two new patterns, namely frequent events and groups of events in data stream. Those new patterns could be treated as a kind of generalization of frequent events, as introduced by Agrawal in [1].

An item set is the subset of items in an analyzed data base. It is called frequent, if it is supported by data more times than a threshold expressed by minSup. The set is frequent regardless its distribution in the analyzed data. The pattern introduced in this paper, in contrast, holds information about time, in which it is frequent. The discovered frequent set in data stream carries not only event's identifier but also the range in which it appears, and its support is above a designated level. With the proposed approach we additionally gain information how events appear within a given period.

The most common approach to data stream analysis consists in discovering event sequences. The sequence is an ordered set of events, which appears in a data stream in a predefined time range. There are many algorithms for discovering this kind of information, in particular AprioriAll [2], GSP [10], PrefixSpan [9]. Another approach to data stream analysis has been described inter alia by Mannila, Toivonen and Klemettinen in [5, 6]. The patterns introduced by them, episodes and episode rules, hold some information about time ranges. An episode is a collection of events which appear in a partially ordered set, in a given time window. An episode rule ascertains that if an episode appears in the data stream in a predetermined time with a probability, another event would appear later. All recalled patterns hold some information about their distribution in the analyzed data stream. This information can be useful in many situations. We try to go further and introduce new patterns in this paper.

A most similar approach to the presented one appears in [3]. The emerging pattern is an item set which supports growth between two analyzed data sets. The emerging pattern carries information that some item set appears rarely in one data set and frequently in another one. The major disadvantage of this method is that the sequences (patterns) could be discovered only by analyzing two data sets. We introduce in this paper new patterns in the data stream, namely frequent events and epoch. Discovering them does not have the disadvantage mentioned above. Below we describe then in more detail.

2. Definitions

Denote E_0 as a set of events identifiers. *Event* is a pair (e, t), where $e_i \in E_0$, and t is a time, where this event occurred.

Data stream is an ordered triple $<t_b^{DS}, t_e^{DS}, S>$, where t_b^{DS} and t_e^{DS} are integer numbers greater than 0, which represent beginning and ending moments of this data stream. S is an ordered sequence of events:

$S=<(e_1, t_1), (e_2, t_2), \ldots (e_n, t_n)>$, and for all pairs $t_b^{DS} \le t_i \le t_e^{DS}$ and $t_i \le t_{i+1}$.

Example 1

Consider sample data stream, presented in Figure 1.

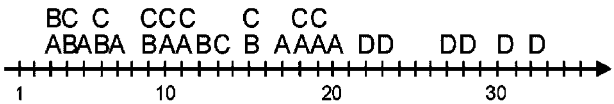


Figure 1. Sample data stream used during all examples in this paper

For this data stream, events identifier set consist of events A, B, C and D, $E_0 = \{A, B, C, D\}$. The data stream has the form <3, 32, S>, where sequence S is presented below:

S = < (A, 3), (B,3), (B, 4), (C, 4), (A, 5), (B ,6), (C, 6), (A, 7), (B, 9), (C, 9), (A, 10), (C, 10), (A, 11), (C, 11), (B, 12), (C, 13), (B ,15), (C, 15), (A, 17), (A, 18), (C, 18), (A, 19), (C, 19), (A, 20), (D, 22), (D, 23), (D, 27), (D, 28), (D, 30), (D, 32) >

This data stream will be used in the next examples.

□

As a *window* in a data stream, we define an ordered triple $W = <t_b^W, t_e^W, WS>$, where t_b^W and t_e^W are integer numbers representing the begin and the end time of the given window, if the conditions $t_b^{DS} \le t_b^W$ and $t_e^W \le t_e^{DS}$ are fulfilled. The subsequence WS contains the all pairs $(e_i, t_i) \in S$, for which $t_b^W \le t_i \le t_e^W$. We define the window length *Wl*(W) as integer calculated as a difference between the end and the begin window time, i.e. $Wl(W) = t^W_e - t^W_b$. For a given event e in window W, we define its support (and denote by *eSup*(W, e)) as the number of pairs (e_i, t_i) in the stream, where $e_i = e$.

We say that *event e is frequent in window W*, when its support *eSup* in given windows is greater than assumed parameter *minESup*, *eSup*(W, e) > *minESup*.

Example 2

Lets consider window W = <20, 28, < (A, 20), (D, 22), (D, 23), (D, 24), (D, 27), (D, 28)>> and parameter *minESup* set by user to value 3. In this window event D appears 5 times, which means that using this data we could say that D is a frequent event in this window.

□

An Interval for integer numbers t_b and t_e such that $t_b < t_e$ we called set:
$[t_b, t_e] = \{x \in N: t_b \leq x \leq t_e \}$.

Number t_b is called *interval begin,* and t_e is called *interval end.* In the following text intervals are denoted with small letters.

Event e is frequent in interval $p = < t_b^P, t_e^P>$, where for each $t \in P$, exists window W, in which e is frequent and where $t_b^W \leq t \leq t_e^W$. Numbers t_b^P and t_e^P are called begin and end time of event frequency in this interval.

Exact interval for event e is interval, in which e is frequent and pairs $(e, t^P{}_b)$ and $(e, t^P{}_e)$ are in S.

Interval p is called *maximal exact interval for event e* when p is exact interval for e and there is no interval q, in which e is frequent and p is subset of q.

Example 3

Assume parameters *Wl* = 8 and *minESup* = 2. Using those parameters we find frequent intervals in the data stream presented in Figure 1. Using those constraints we can say that D is frequent in interval p_1 = <26,34> because there is a window W = <24, 32,{(D, 24), (D, 27), (D, 28), (D, 30), (D,32)}> in which D is frequent and each $t \in p_1$ encloses in this window.

p_1 is not an exact interval, because in analyzed data stream there is no pairs (D, 26) and (D, 34) which meets the case of event at begin and end time of interval. As an example of an exact interval where D is frequent we can show interval p_2 = <27, 32>.

p_2 is not a maximal exact interval for event D, because we could show other interval p_3 = <22, 32>, which include p_2. This interval is maximal, because in the presented data stream we could not show other, larger interval which begins and ends with event D.

□

Frequent event in data stream is such $e \in E_0$, which in given parameters window length (*Wl*) and minimal support (*minESup*) is frequent at least in one window. Existence of such a window implies one interval in which e is frequent.

Length of the interval A, denoted *Len*(A) we called number which is calculated as subtraction begin of interval from end of interval.

We say that *two intervals A and B are similar*, when they have common part containing at least one element. In other words, which common part length is greater than 0.

Similarity factor of two intervals A and B, we called parameter SF calculated using presented below formula:

$$SF(A,B)=\begin{cases}\frac{Len(A\cap B)}{Len(B)} & \text{when } A\cap B \neq \emptyset \text{ i } Len(A) < Len(B)\\ \frac{Len(A\cap B)}{Len(A)} & \text{when } A\cap B \neq \emptyset \text{ i } Len(B) < Len(A)\\ 0 & \text{in other cases}\end{cases}$$

An appearance of the event e, we called the set of all maximal exact intervals in each event e is frequent, using fixed parameters *Wl* and *minESup*.

During work with real data useful could be ability to filter shortest intervals. In this case parameter *minFEP*(minimal Frequent Event Period) could be used. When it is used, only intervals that have length greater than *minFEP* are added to the appearance.

An appearance P for a given event e_1 covers an appearance Q for given event e_2 using SF factor, when for each interval belonging to the set Q, could be found similar interval in the set P, which calculated SF factor is greaten then given parameter *SFmin*.

We say that *appearance P and Q are similar*, when using given *SFmin* parameter P covers Q or Q covers P.

Group of the events we called set of the events, where using the fixed *SFmin* parameter all pairs of the events in it have similar appearances.

Appearance of the group of events we called appearance which is common part of appearance of the all member's events of the group.

For discovery of the groups could be added several conditions. Those conditions allow discovering only important to the analyst patterns. During described later experiments two such parameters are used – *grSumMin* and *grSumMax*. Those two parameters are conditions that allow only discovery of groups which summary length of all periods in appearance have length greater than parameter *grSumMin* and not greater than *grSumMax*.

3. Experiments

The main purpose of conducted experiments is to confirm two theses. The first is that the above mentioned method could lower the number of discovered patterns in opposite to all previous methods. The second is that discovered patterns contain new and useful knowledge.

The first part of this section describes used data. In all experiments only real, not synthetic data sets are used. In the second part results of conducted experiments are described. In this section quantitative result are presented and discussed. For example, numbers of events in initial data or numbers of proposed discovered patterns. The last part describes knowledge which could be obtained from analysis of computer logs. This is the next step after Data Mining discovery in KDD, which analyzes and use knowledge obtained in discovered patterns.

3.1 Description of used data sets

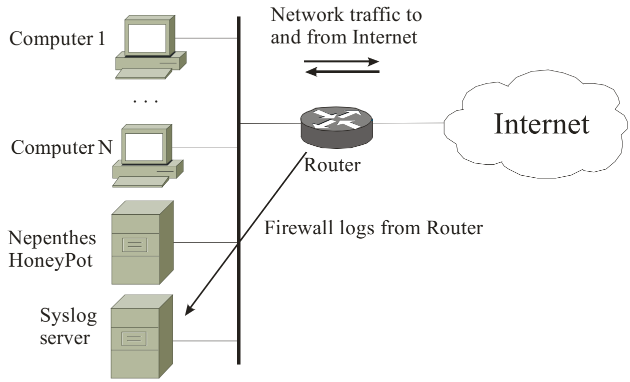


Figure 2. Topology of the network in which logs are taken and where the HoneyPot system was placed

Experiments are conducted on three types of data sets. As the first one, data set logs from firewall are used. Logs are collected from a home network during a normal user's activity. Additionally, an activity of the HoneyPot system, which works every few times for few hours in this network, is recorded. Topology of the network, from where logs are acquired, is presented in Figure 2. Traffic travelling to or from the Internet was forwarded by router. For carrying the experiments two kinds of events are worth noticing: the drop of packet that does not fulfil a security policy and the start of new TCP connection. Information about these events was logged and sent via the syslog protocol to the server. Received and stored on it data was then analyzed. This data set is called syslog.

The second type of data used during experiments is logs from Nepenthes [7] - a low interaction HoneyPot. This program simulates vulnerable services and catches malicious code (shortly malware), which attacks it. Consequently, the whole hostile activity of malware, especially IP addresses, names of malware and sometimes even some passwords and logins to hacker's machines, are logged. This information is used as an input data for presented methods. During experiments two data sets of this kind are used, Nepenthes2006 and Nepenthes2008. Because both data sets give almost identical result, only recent is described later in more details. The HoneyPot system worked in the same network, which is presented in Figure 2.

The last type of analyzed data is an electronic mail. In experiments authors' emails are analyzed. The data set is called Cisco and contains all emails from students and staff in Cisco Academy courses. In data set there are 1020 emails received from September 2006 to August 2007. In presented experiments only information about a sender is used.

3.2 Quantitative results

Experiments prove that the proposed method significantly reduces number of obtained patterns. This is an enormous advantage, because after an automatic Data Mining step, analysts have to consider discovered patterns. When the method discovers fewer meaningful patterns, person who analyzes results could focus only on the area of interest. Hence, his work could be more effective. Table 1 presents results of conducted experiments. At rows 3 and 4 all used parameters are shown. Parameters are set by

authors during an initial step, when general character of data sets is found. This process takes advantage of field's knowledge. Presented parameters are optimal for those data sets, and lead to discovery of very interesting and meaningful patterns.

If we compare in Table 1, row 2 (number of events in data set) with rows 4 and 8 (number of discovered patterns, respectively frequent events and groups of events), significant reduction of number could be noticed. Analyzing of such number of initial events (presented in row 2) without any automatic techniques and discovery from them any knowledge is almost impossible. After use of one from proposed method, number of discovered patterns is reduced that knowledge discovery and use of discovered patterns is possible (numbers presented in rows 7 and 8). Please note that in rows 7 and 8 there are difference between presented numbers. This change appears due to some discovered patterns are not presented by software to the analyst. Removed patterns are subsets of those not presented, and they contain only incomplete, not useful knowledge.

In the Table 1 results of only one experiment execution per data set are presented, those that are discussed in more details in the next paragraph. Proposed method gives fewer patterns not only using those discovery parameters. Below in Figure 3 to Figure 5 are shown results of conducted experiments using various values of *minESup* parameter. In all cases number of discovered frequent events in data stream is smaller then using frequent events. For 2 of 3 data sets number of obtained frequent interval is smaller as well. In the Nepenthes2008 data set number of frequent intervals is greater then number of frequent events because of some sampling used during HoneyPot system execution.

Table 1. Quantitative results of conducted experiments

1. Data set name	**Syslog**	**Nepenthes 2008**	**Cisco**
2. Number of events (different)	254 619 (2880)	49 230 (5635)	1020 (185)
3. Frequent Events discovery parameters	minFEP 1 day eSup 3 Wl 3 days	minFEP 15 min. eSup 3 Wl 1 hour	minFEP 1 day eSup 2 Wl 3 moths
4. Number of discovered frequent events (frequent intervals)	314 (498)	930 (1907)	94 (97)
5. Groups discovery parameters	SF 0.90 grSumMin 7 days grSumMax 20 days	SF 0.9 grSumMin 5 hours grSumMax 45 days	SF 0.75 grSumMin 2 days grSumMax 6 months
6. Number of discovered groups (fixed parameter SF = 0)	15139	4531	> 500 000
7. Number of all discovered groups	82	109	1523
8. Number of groups when subsets of groups are removed	40	25	38
9. Number of groups which size is greater than one	7	9	16

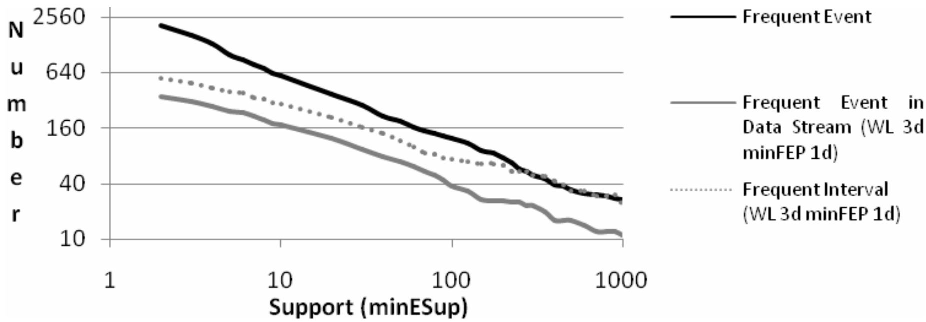


Figure 3. Number of discovered patterns in the Syslog data set

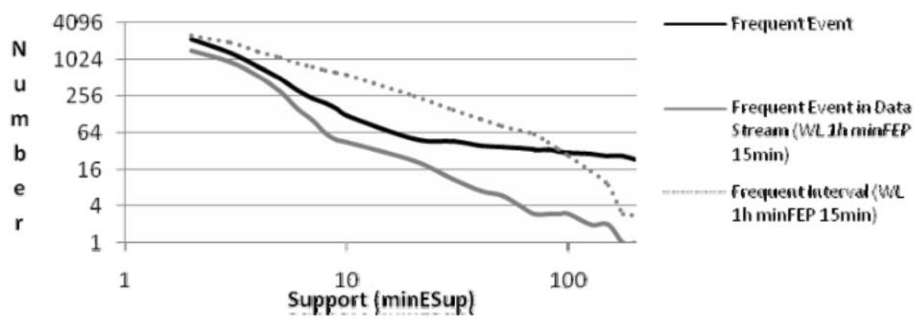


Figure 4. Number of discovered patterns in the Nepenthes2008 data set

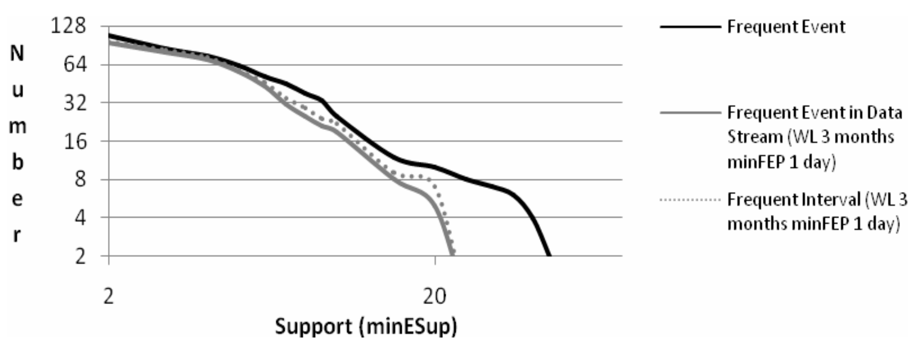


Figure 5. Number of discovered patterns in the Cisco data set

3.3 Learned knowledge from discovered patterns

Table 1, presented in the previous section, shows a significant reduction in number of initial events in reference to a number of discovered patterns. However, for Data

Mining algorithms not only the number of discovered patterns is important; knowledge which they contain is even more important. Only meaningful patterns, those who represent real useful knowledge, could be used for analysis of data. In this phase of KDD process expert's help is needed. In this section, examples of patterns that carry knowledge that could be automatically discovered and later use are discussed.

Experiment conducted using the syslog data set simulates situation when computer security expert have to obtain from delivered logs as many as possible information, which could be used in investigation – legal or inside institution after some security break. The work should give as many as possible information about people and machines which activity was recorded in obtained data. This kind of process is called Computer Forensic, and because of the popularity of computers, it is more and more often applied. In the first step automatically discovered events are verified by expert. Using information from logs, like IP address and destination, TCP port etc., picture of user's activity could be created. In this data set not only general user's activity could be discovered. The majority of generated from observed network traffic is connected with web sites. Those names and probably reason of visit could be explained, when logged IP address are changed to human readable form using DNS reverse lookup. All next deductions are made using pre-process in this manner address. In next section sample patterns and its explanation are presented in more details. Several addresses in discovered frequent events are connected with update websites for some programs (Zone Alarm, Microsoft Windows, and Mozilla). Using that information, knowledge about used software could be gained. Looking at other addresses, could be obtained information about favourite newspapers in the Internet (BBC, Gazeta.pl, Rzeczpospolita, Dziennik) and entertainment portals (o2.pl, YouTube, Wrzuta). In analyzed data set patterns, which suggest that computer is used for communication, are discovered. This class contains frequent events representing traffic connected with GaduGadu communicator and two e-mail portals (poczta.o2.pl and gmail). A user which data was analyzed uses the Internet for e-banking activity. In discovered patterns there are few connected with two bank web sites. Secure ciphered access, using https protocol, proves that some bank orders are executed – and user really use e-banking features.

Previously it was not possible to obtain any special information about user's interests; it gave only general behaviour of average user of the Internet. But other patterns could give some more specific information. Firstly, in data was found patterns that joined person, whose activity was recorded in logs with two educational institutions. There are few patterns that represent usage of machines belonging to Faculty of Electronics and Information Technology. Those patterns suggest that someone logged into few machines in this organization (ssh service), use mail services (smtp and ssmtp service) and use some inside portal used for managing students grades (https). The second institution is connected with Cisco Network Academy, and patterns that prove this fact, represents connections to restricted e-learning system. Second, unusual group of patterns reveals information about other interest of user; there are few discovered patterns that suggest this interest. All of them are connected with access to financial web sites. In this group could be mentioned web sites of Warsaw Stock Exchange (www.gpw.com), financial branch of one News portal (giełda.wp.pl) and home pages of three Investment Funds (SEB, PKO TFI, Legg Mason).

All described before information is based only on frequent events; however, some interesting knowledge could be concluding from the second pattern – groups of events. In this case obtained knowledge is even more interesting. In conducted experiments on

the syslog data set two groups are spectacular. First group consists of four events. Three of them have been already discussed, and are connected with a three Investment Funds. But together with them forth pattern represents web access to some address of company named Instadia. Before, when frequent events patterns are analyzed, it was not clear why this address appears in discovered patterns. Whereas, using information about probably connection with one of those Investment Funds quickly reveals interesting information. The Instadia Company is an owner of the ClientStep program, which is used by SEB Investment Fund (http://www.webanalysts.info/webanalytics/do-others-use-web-analytics/). The second very interesting group is connected with the HoneyPot activity. In this group three of events are connected with easy understandable hostile traffic to the HoneyPot (connections from the Internet, to well known in security world TCP ports 135, 139 and 445). Previously unknown pattern represents connection to IP address 72.8.XX.XX and port 31, in combination with other members of this group became explained. Because pattern represents a connection from the HoneyPot system and was recorded many times, with high probability, we could assume that from this address some malware is downloaded after infection. This information is acknowledged in the HoneyPot logs (we return to this pattern in description of the next experiments data sets).

Additional knowledge in these patterns is contained in frequency periods, periods of time when discovered patterns or groups of patterns are frequent. For example, in the syslog data set the analyst could observe an interesting behaviour. Almost all interesting and specific for this data sets both events and groups, like HoneyPot activity, access to educational sites, at the same time, Investments Funds web pages are not frequent in similar periods. There are two gaps, one between 17 to 22 February and second between 5 to 10 March. Both of them suggest a break in activity. This information could be interesting, for example, if such an analysis is performed in some investigation. This could be an important fact suggesting that a suspected person was, for example, on holidays during that time. In this case the author, person whose data are analyzed confirmed that in this period he go to some training and then to holidays.

The Nepenthes2008 data set give valuable knowledge as well. Discovered patterns, in this case only groups of events, simplify identification of the malware type and other information connected with it. The first important, discovered fact is connected with malware names. Sometimes file names of malicious code are randomly generated. But, when the same name is used, it could be easily discovered by proposed method. Using provided file names of suspected code; other machines could be checked. In this case we could treat presented system as an automatic generator of anti virus signatures. In the Nepenthes data set during experiments 18 file names are discovered, for example: WindowsUpdater.exe, antiv.exe, iPodFixer.exe or kwjwjshshsx3.exe and moo4.exe. But, proposed method gives more valuable knowledge. Depends on the type of malware, additional information is contained in discovered patterns: IP addresses of machines from which malware is downloaded, used protocol and its specific ports, or even authentication data such as logins or passwords. Knowledge discovered in patterns generated from HoneyPot logs could be easily used for protection of other networks. Most valuable information is stored in IP address. When the address is discovered in a pattern, it could be assumed with the high degree of probability that the hostile machine could be used by attacker. In this case the address should be blocked on firewall, protecting other machines. Apart from this step, other legal actions could be performed, for example, sending information to the owner of machine, because in

many cases he is not aware about infection. Like in previous example, such situation could be treated as an automatic Intrusion Detection System (IDS) signatures generator.

In this experiment, other advantage of proposed patterns could be easily presented. Simplicity of generated patterns is one of them. Below a sample excerpt of discovered patter is presented.

```
user=reviv
password=hxedb0x42
address=ssffttpp.jackill07.biz
port=31
file=msv.exe

From 2008-02-09 18:47:45 to 2008-02-09 20:08:27 01:20:42
From 2008-02-10 14:22:48 to 2008-02-10 16:58:33 02:35:45
From 2008-02-11 15:32:10 to 2008-02-11 16:09:49 00:37:39
. . .
```

Verifying this excerpt, all useful information could be easily understood by the network expert. In the pattern we have all information about specific malware, in this case contained in file msv.exe. Other information like the address (in a pattern shown using human readable DNS name), a port and authentication data are presented as well. At the bottom of the pattern there is information about intervals when a group of events frequently appears - in the presented excerpt time intervals when the HoneyPot system is working. What is also interesting, when domain name, from which malware is downloaded, is changed to IP address we receive 72.8.XX.YY address – this same discovered in the syslog data set.

Like previously data set information about time gives valuable knowledge as well. For example, in the data set Nepenthes2008 are discovered several patterns which have the same IP address 220.95.XX.YY. All patterns have frequency periods that do not cover each other. Furthermore, file names which were downloaded from this address change only in one character: moo.exe, moo2.exe, moo4.exe etc. Similar changes could be observed when an address has a human readable form, using DNS. This case is presented in sample excerpt - this address during time when data was collected has changed to several addresses in 72.8.XX.00/16 network.

In the last data set called Cisco, the proposed method is used for grouping e-mail's senders. In automatically discovered 14 groups, 6 of them have more than 5 members, and only those are examined. In this experiment each member represents one sender, and discovered patterns represent some connections between senders. The system assigned senders to a group according to similarity in its behaviour collected in test data set. It occurred that those groups are almost identical with the actual groups. To prove usability of propose method, automatically generated groups have to be compared with real course lists. After this step, the result is promising. Three of those groups consist only of instructors and people who attend to the same course groups. In the biggest group except instructors there are 9 people from 15 that begin a course. What should be highlighted all of them successfully ended course. The similar situation appeared in next groups – discovered 10 people from starting 16. In this group 9 people ended course on time. In other discovered patterns grouped senders attended to two groups. For this result influence has fact, that those two course groups for some period of time are trained together.

4. Conclusions

Most of currently used patterns in Data Mining treat data sets globally. Only patterns that are enough interesting in the whole data set are discovered. Information about local changes in pattern distribution is lost. In the paper novel patterns are presented, frequent events and groups of events in data stream. Both of them could discover local changes in analyzed data sets.

Several conducted experiments prove that using presented method some new information could be obtained. In the analysis of firewall logs detailed behaviour of user could be presented. Additionally some firstly not seen connections between some of address could be revealed. In the Nepenthes logs detailed information about the malicious code could be obtained, and easily used in AV or IDS systems. The last experiment shows that this method could discover, in stored mail, groups of people, whose communication in some period of time was intensive. This information could be especially useful in computer forensic.

What have to be noticed using those methods, the number of discovered patterns is reduced in contrast with other, before used techniques. This is great advantage of the described method, which allows analysts focus on discovered knowledge without verifying all data.

Acknowledges

This Research has been supported by grant No 3 T11C 002 29 received from Polish Ministry of Education and Science. The author thanks Professor Marzena Kryszkiewicz for her valuable comments especially during mathematical formulation of introduced patterns, PhD Jacek Wytrębowicz and PhD Krzysztof Szczypiorski for many inspiring talks, and my friend Ania for patience during punctuation and language corrections.

References

[1] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, In Proc. 1994 *Int. Conf. Very Large Data Bases (VLDB'94)*, 487–499, Santiago, Chile, Sept. 1994.

[2] R. Agrawal, R. Srikant, Mining sequential patterns, In Proc. 1995 *Int. Conf. Data Engineering (ICDE'95)*, 3–14, Taipei, Taiwan, Mar. 1995.

[3] G. Dong, J. Li, Efficient mining of emerging patterns: Discovering trends and differences, In: *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (SIGKDD'99)*, 43–52, San Diego, USA, 1999.

[4] K. Julish, *Data Mining for Intrusion Detection, Advances In Information Security*, Kluwer Academic Press, 2002, 33-62.

[5] M. Klemettinen, *A Knowledge Discovery Methods for Telecommunication Network Alarm Database*, PhD. Thesis, University of Helsinki, January 99.

[6] H. Mannila, H. Toivonen, and A.I. Verkamo: Discovering frequent episodes in sequence. In: *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*. Montreal, Quebec, 1995, pages 144–155.

[7] Nepenthes HoneyPot, http://nepenthes.mwcollect.org/.

[8] S. Noel, D. Wijeesekera, C. Youman, *Modern Intrusion Detection, Data Mining, and degrees of attack guilty, Advances In Information Security*, Kluwer Academic Press, 2002, 1-31.

[9] J. Pei, J. Han, at al., PrefixSpan: Mining Sequential Patterns Efficiently by Prefix Projected Pattern Growth, In. Proc. *Int. Conf. Data Engineering (ICDE'01)*, 215-224, Heidelberg, Germany, April 2001.

[10] R. Srikant, R. Agrawal, Mining sequential patterns: Generalizations and performance improvements, In Proc. *5th Int. Conf. Extending Database Technology (EDBT'96)*, 3-17, Avignon, France, Mar. 1996.

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-81

Mining Local Buffer Data

Andrzej SIEMIŃSKI
Wrocław University of Technology, Institute of Applied Informatics,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
e-mail: andrzej.sieminski@pwr.wroc.pl

Abstract. Web mining employs the techniques of data mining to extract information from the Web for a variety of purposes. The usual sources of data are the log files of WWW or proxy servers. The paper examines the possibility of using the local browser buffer for that purpose. The data that could be extracted from both types of logs are compared. It turns out, that despite its limitations the browser buffer is a rich source of unique data about user navigational habits and the properties of the fragment of the WWW that he/she visits. The cache contains the both the full body of a WWW object as well as the header control data sent by the server. Additionally the cache includes some basic information about the usage pattern of each object. Therefore it is possible to study the susceptibility to buffering the objects which is measured by the CF (cacheability factor) and to study the word diversity of Internet texts seen by the user. The CF factor provides an objective measure of the web site caching potential and thus makes it possible to infer about latency of the web site. The word diversity study tests the compliance of the Internet texts with the well known Zipf and Heaps Laws' that are valid for all natural languages. That part study could be used for the optimization indexing engines or the recommendation of pages potentially interesting for the user.

Keywords. Internet, data mining, behavioural targeting, browser cache, latency, Zipf law, Heaps law, cacheability

1. Introduction

Web data mining is a now the focus of study for an increasing number of researchers. It is the study of data mining techniques to automatically discover and extract information from the Web. The scope of applications of the research in this area ranges from improving the quality of service by reducing the user perceived latency to personalizing the Web content by generating user specific pages or supplying him/her potentially useful links [1]. Web data mining is subdivided into sub areas: Web formal features, Web content analysis and Web usage patterns.

The most important issue in data mining is starting with the right data. The logs collected at: Web servers, proxy servers or in a local network are the usual sources of data for Web mining. All of them have their limitations pertaining availability and scope of contained data. The paper proposes an alternative source of data - the local system cache of Internet objects. The cache contains all recently requested objects and header control data supplied by the WWW server. It forms therefore a full picture of user navigational habits and is a complete copy of the fragment of the WWW that he/she recently visited. The described study is user centric. Its aim is provide data that could improve the operation not of the Web server but of the local workstation. As

opposed to previous attempts to utilize local buffer data that had concentrated only upon the detection of usage patterns [2] the study attempts to utilize this unique and a rich source of data in all three mentioned above areas:

- the evaluation of the scope and nature of the cacheability of Web objects (Web formal properties);
- the repeatability of user behaviour (Web usage pattern);
- the word diversity (Web content mining).

The rest of the paper is organized as follows:

Section 2 compares the data that can be extracted from the different types of logs with the data from the local cache. Section 3 describes the operation and specifies the content of the local buffer. The repeatability of user behaviour is discussed in section 4. Section 5 introduces the CF (cacheability factor) and discusses its interpretation. The compliance of texts extracted from the cache to the Zipf and Heaps lows is presented in the Section 6. Section 7 concludes and outlines future research areas.

2. Sources of Data for Web Mining

The sheer size of Internet makes it impossible to analyze all of its objects. Therefore it is so important to start with the right data. The ideal set of data should be easy available, complete, span over a reasonable long period of time and be in many cases user specific. The usual source of data for Web mining are logs collected at different levels of Internet infrastructure: Web server, proxy cache and local network. All of them have their advantageous and disadvantageous.

The Web server logs are collected at each Web server but as they contain information widely considered to be sensitive there are not many such logs that are free for research purposes. Popular logs such as Music Machine log [3] or World Cup log [4] are several years old. The log does not contain the body of the requested data; hence Web content mining is not possible. The data is well suited for the analysis of one Web site usage preferences of groups of users. The analysis starts with the complex task of identification individual user sessions within a log. The problem is far from being trivial as e.g. the proxies filter out many popular objects and substitute user's IP numbers. This could easily lead to the merging of sessions of different users. Even more challenging is the collection of user sessions over a longer period of time. There is no general, unique identifier, as the majority of users do not have a fixed IP numbers because they are dynamically generated by Internet providers. What is more important the users often use proxy caches or simply access the Internet from different environments (e.g. work and home). The only reliable way to identify a user is to require login to a server. This procedure is not accepted by many users as being bothersome or because they are simply too much concerned about their privacy. Most users prefer to stay anonymous in the Web and would and do not wish to make public their Web access data.

The data is limited to a single server so identifying all of user preferences is impossible. This makes the behavioural targeting not possible at all. It should be stressed that lower level caches satisfy requests to many popular objects so it is impossible to reconstruct a complete user session. This is not crucial when we study the

visited HTML objects. They are usually loaded from the server on every access. The feature hampers significantly the measurement of the cacheability of objects.

The availability of proxy logs is far greater Web server's logs. Large proxy log files are available for each day and are free for research purposes e.g. the popular logs from the IRCache projects [5] that were used by so many researchers. The log formats are normalized and even programs for the statistical analysis of popular log formats are available [6], [7]. However that type of data has also serious disadvantages. The logs of proxy servers are not available for every region and they contain only basic header data. The requested objects are not available and the direct study of Web content is not possible. In some cases even the download time is not available. The data is anomnized to protect user privacy, so the collecting of multi session single user data is impossible. The user session identification is almost as challenging as in the case of Web servers. The log data removes only one limitation of the Web Server logs: they cover a wide spectrum of servers.

The local network logs are even scarcer than the Web server logs. There are only a few publicly available log data of that type but they are generally outdated as e.g. the Boston University Log [8] collected more than ten years ago. The identification of a user during one session does not pose any serious problem but due to the anominization process the multi session information is still not available. The scope of the visited Web sites is complete but the data is not detailed enough: the log contains only the most rudimental information such as the requested URL or download time.

As stated above one disadvantage common to all the discussed logs is the inability to identify user data over a longer period of time. Precisely that type data is necessary for the behaviour targeting. The behaviour targeting is a relatively new and promising research topic in the area of Internet recommendation. In the regular recommendation system it is assumed that what pages Web site visitors click on and where they go from those pages indicates at least a presumptive interest in buying products related to the topics that they click. In the behaviour targeting paradigm the importance of gathering data from different Web servers over a long period of time is emphasized. The input data must therefore meet very strict requirements:

- the user must be uniquely identified;
- the data must cover whole user interaction with the Web;
- the data must be collected over a reasonably long period of time.

In practice the collection of such data requires user collaboration. The usefulness of such data is appreciated not in research but also in the industry. Software companies can offer incentives to Web users prompting them to make reveal the way interact with the Web. It is estimated that the proportion of users are willing to cooperate exceeds slightly 20%. Such a permission-based cooperation exists e.g. in the GAIN Network [9] but the constant monitoring the user behaviour and sending data over network provokes serious privacy concerns. The legislative and regulatory framework that governs behavioural marketing is described in [10]. The participants of the GAIN Network are "paid" for making available their preferences by distributing a variety of popular software applications.

Similar service is offered by Google. The Google Desktop offers to search the local files using the well known capabilities of Google which makes the search private files and the Internet alike. There is a price for the convenience. The "advanced features option" sends the pattern of user behaviour to the company to enable behavioural targeting. The feature, which could not be entirely disabled, had raised serious privacy concerns [11].

The data gathered by GAIN network or other similar companies is the property of commercial companies and therefore is not available for research purposes.

3. Local Buffer Data

Some of the mentioned above problems could be solved by collecting data about user interaction with the Web directly at their source of origin - at the browser cache. It collects data locally, so as long as the data is processed locally there are no privacy concerns. The buffer is not a log, it contains a set of objects and therefore it does not allow to e.g. to produce a sequence of visited pages. The disadvantage is matched by the completeness of data - all pages are to be found, no matter on which server were accessed. Web mining procedures require intensive processing. Analyzing data locally disperses the processing and therefore it could be more computational intensive and thus hopefully deeper. The local collecting and processing of data neither does not allow us to use the buffer data for the content modification. This mode of operation requires sending "distilled" data about user preferences to the WEB server e.g. in a form of an introductory cookie. A server sensitive to such data could process the cookie, assign the user to a proper group and then piggyback its hints or generate a personalized page in the usual way. The advantage of the solution is the clear separation of responsibilities. The workstation processes data specific to a particular user whereas the server processes data on users groups. The disadvantage is that the solution requires a definition of language to describe user preferences and supported both at the local workstation and server. The discussion of such problems goes far beyond the scope of the paper.

In what follows the most popular Windows operating system is considered. For each user the system maintains a local buffer - the TIF (Temporary Internet Files) which is a simple data base. The database contains a table with all recently requested files and their attributes. It does not provide any log information. The TIF operates in one of 4 available modes. In the most popular mode "automatic refresh" mode the system automatically inserts and deletes objects from the TIF The standard WINint library contains several functions to process the TIF. The way the buffer operates is unfortunately not publicly available. A detailed analysis of the TIF operates had revealed the following rules:

- all requested WWW objects are fetched from the buffer;
- the total size of objects in the buffer could not exceed 70% of its capacity;
- the objects that were deleted have: low Hit Rate (usually 1), were not recently accessed, the WWW server had not sent cache controlling data.

As a result the TIF size of at least 100 Mb guarantees that fast all files from last the last several days are preserved.

The TIF buffer contains elements of three types: files, cookies, and URL history. In the study only the files and the URL history were analyzed. During the study following attributes were used:

1. URL;
2. Object size;
3. Expire, Last-Modified, Last-Access which are self-explanatory timestamps;
4. Entry-Type to distinguish the files from the cookies and history entries;

5. Header that contains some useful information such as file type or caching history;
6. Hit-Rate that could be used to account for the popularity effect;
7. Ext that contains the full body of Internet object.

Only Hit-Rate and Last-Access attributes are maintained locally, the rest of attributes originates at the WWW server.

The files were further divided into 4 categories using the ContentType header included in the Header attribute. The categories were:

- application (APP),
- Image (IMG),
- Text (TXT), and
- Other (OTH).

The last category was used when the ContentType header was not supplied by the server.

During the experiment two data sets were analyzed. The first (LAB) consists of data collected from about 10 workstations at the University Laboratory and the second (IND) includes data from 2 workstations of the University staff. The laboratory computers were used by a number of students, all of them logged to the system using the same name so their interaction with the Web was kept in a single file.

The Web pages are accessed not in a uniform manner some pages are loaded more often than other. Therefore many studies have taken into account the so called page popularity effect. The LABp and INDp refer to modified data sets in which the each entry was assigned a weight equal to value of the Hit Rate attribute. This enables us to take into account the popularity effect in manner that is not perfect but is certainly acceptable.

Each selected workstation had the TIF with entries that were at least 2 months old. The number of all entries was in the individual TIF's were in the range from 3,000 to 20,000. The total number of entries was nearly equal to 120,000. The files make up 106,800 entries.

4. Repeatability of User Behaviour

The URL's of Web pages could be divided into two groups: dynamic and static. The dynamic pages are generated by the WWW server on each request and their content could be personalized. The distinction is important, as e.g. the dynamic pages are less sustainable to caching. The usual way to distinguish between the two types of pages is to check for the occurrence of specific strings within the URL field: "?", "=", ' (a query to a database system) and "cgi", "cgi-bin" or .asp (calling a Common Gate Interface or Active Server Pages Script respectively). The average Hit Rate for the two types of pages and for all data sets is shown in the Table 1. The results were early reported in [12].

The LAB data set contains more ad hock generated URL's that are product of various search engines. The IND data set is far more static URL's and the stability of interest is even more apparent when the popularity factor is taken into account. The Average Hit Rate for dynamic pages is similar in both data sets but the Individual User shows once more again far greater stability of interest.

Table 1. The Share and hitrate of static and dynamic Web pages [12]

Data set	%Occurences		Avg(HitRate)	
	Dynamic	Static	Dynamic	Static
IND	30	70	2.39	5.87
INDp	15	85		
LAB	52	48	2,27	3.2
LABp	43	57		

It should be stressed that the above modifiers "static" and "dynamic" apply to the pages URL's and not their content and are therefore meaningful in the context of prefetching and not caching. The static URL belong e.g. to news Websites whereas the dynamic are typical for shops or search engines. As a matter of fact the content of many pages with static URL's changes quite often. The formal changeability of Web objects is discussed in the next section. A page with a static URL could not be safely read from a cache but it is safe to assume that it should be prefeched.

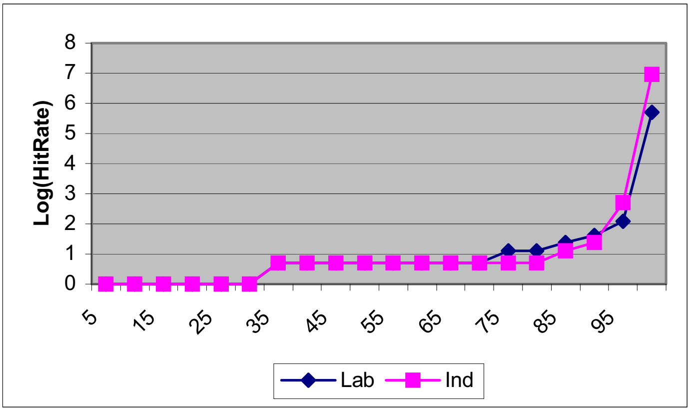


Figure 1. The distributions of Hit Rates for pages with static URLs

In both data sets about one third of all URLs were accesses only once and in the Ind data set over 80% were accesses less than 3 times. The maximal values of Hit Rate were equal to 302 and 1049 for the Lab and Indy Datasets respectively. The users visit on regular basis only a very limited number of Internets sites.

A Web site consists of the so called navigational pages and content pages. The 5% of most popular URLs are mostly the navigational pages. The navigational pages have stable or predictable URLs, see examples on the Figure 2. The navigational type URL's are written using italic font. Pages with such URLs are stable over a considerable period of time. In the Case B the structure of URL is slightly more flexible but it still predictable. The middle part of the URL (written in bold) changes on daily basis but the pattern is clear to see: the string "wydanie" (issue in English) is fixed and the rest indicates the date of the issue.

The buffer data clearly indicates that a relatively small proportion of URL's are visited regularly and that their content is stable or predictable. Therefore they are good candidates for prefetching.

Case A:
http://www.gazeta.pl/0_0.html'
http://serwisy.gazeta.pl/swiat/0_0.html'
http://wiadomosci.gazeta.pl/wiadomosci/1_53600_2418753.html'
Case B:
http://www.rzeczpospolita.pl'
http://www.rzeczpospolita.pl/gazeta/wydanie_041215/index.html'
http://www.rzeczpospolita.pl/gazeta/wydanie_041201/kraj/index.html'

Figure 2. The two typical forms of static Urls taken from news Websites

5. Cacheability of Web Sites

The HTTP protocol clearly specifies whether an object could be cached or not [13]. A thorough study of reasons of objects uncacheability is given in [14]. The study into the scope and nature of WWW objects cacheability is vital for both caching/prefetching and the indexing of Web content by search engines [15].
The majority of the uncached objects is dynamically generated. They are received with an attached cookie or are assigned uncacheable response codes or header fields by the WWW server. There were many attempts to estimate the scope of cacheable objects. The attempts span over a period of several years and use different methodology so the results differ to some extend. Early results on the scope of cacheable objects are reported e.g. in. [16] or [17] and are based on the analysis of proxy log files.
The TIF does not give us any direct information that could be used to identify the uncacheable objects. The header attribute contains however data about cache hits and misses that occurred while the object was fetched. Potentially cacheable objects include inside the Header field strings like "Cache HIT" or "Cache LOOKUP". The percentage of objects containing such strings is shown in the Table 2.

Table 2. Percentage of cacheable WWW objects

Object Type	Without popularity	With popularity
APP	59 %	52 %
IMG	96 %	99 %
TXT	63 %	69 %

The column "Without popularity" shows the percentage of cacheable objects of a given type whereas the "With popularity" column weights the previous result using the Hit_Rate attribute. The results are higher (especially for IMG files) than previously reported. This is probably due to the TIF replacement algorithm which does not remove elements with set Expire_date attribute witch is much more popular for IMG objects than for other objects types. Another justification for the discrepancy in results is that the persistent IMG files are fetched mostly from the local cache so they do not occur

that often on proxy level. It was also observed that the APP type objects (such as JavaScript programs) are more easily cached on less popular pages.

The mere statement of object cacheability is not enough to estimate its caching potential. To estimate it, the objects' freshness life span should be known. An object with longer freshens life span are more likely to be fetched from a buffer than from the original WWW server. In the experiment the Expires field of the TIF was used to calculate the freshness life span. This spares us considerable processing that is required to calculate it from the header data supplied by a WWW server.

In the TIF data the Expires defines the EL - the exact freshness lifespan of an object. Unfortunately the field is set for less than 30% of all objects. The figure corresponds well with previously reported results. The inability to calculate the freshness life span could hamper significantly the performance of a cache. Therefore the HTTP protocol allows the proxy caches to use a heuristic algorithm to estimate the Expire date. The protocol does not introduce any specific algorithm; it describes only its properties. Proxy caches employ usually an algorithm that uses the far more popular Last_Modified field. The PL (predicted lifespan) is set to:

- EL, when Expire date is provided by the server;
- p*(Last_Access-LastModified), when EL is not set and Last_Modified attribute is provided by the server
- 0 in all other cases.

The usual value of p is 0.2.

To measure the susceptibility to caching the CF (cacheability factor) was introduced [17]. The factor is a floating point number from the range [0..1] and it could be interpreted as a probability that an object would be loaded from a buffer and not from a WWW server. Objects with CF = 0 could not be cached at all as each request requires a download form a server, whereas an object with CF=1 could be always loaded from a cache.

The susceptibility to caching does not depend on objects' PL alone but also on R - the refresh rate that is the time span that separates consecutive requests for a given object. Not all values of R are equally useful. The values should refer to typical user behaviour. The values from the Table 3 are used in experiments.

Table 3. Selected values of the refresh rate

Description	Code	Interpretation
1 minute	RMin	Skimming through a Web site
10 minutes	R10min	Carefully reading pages
1 hour	RHour	Return to a page during the same session.
1 day	RDay	Return to a page on daily basis.
1 week	RWeek	Regular but not frequent returning to a page.
1 month	RMonth	Sporadic return to a page.

Cacheability Factor CF(x,R) for object x with respect to refresh rate R is then defined as follows:

$$CF(x,R)= \frac{(PLn(x,R) \quad - \quad R)}{PLn(x,R)}$$

where:
R is the refresh rate of the object x (in seconds).

PLn(x,R) is the normalized predicted lifespan which is defined as follows:

$$PLn(x,R) = \begin{cases} PL(x) \;\; for \;\; PL(x) > R \\ R \qquad\quad otherwize \end{cases}$$

The values of CF of different object types for the selected Refresh rates are shown in the Tables 4 and 5, respectively [17].

Table 4. The CF values of the refresh rate for the selected file types, the popularity effect not accounted for

File Type	Rmin	R10Min	Rhour	Rday	Rweek	Rmonth
APP	0.30	0.29	0.27	0.20	0.08	0.03
IMG	0.78	0.78	0.75	0.64	0.42	0.22
TXT	0.28	0.28	0.24	0.18	0.11	0.05
ALL	0.64	0.63	0.60	0.51	0.33	0.17

Table 5. The CF values of the refresh rate for the selected file types with the popularity effect

File Type	Rmin	R10Min	Rhour	Rday	Rweek	Rmonth
APP	0.56	0.51	0.46	0.35	0.18	0.07
IMG	0.95	0.94	0.93	0.82	0.61	0.38
TXT	0.50	0.48	0.45	0.35	0.19	0.09
ALL	0.87	0.86	0.84	0.73	0.54	0.33

The local cache contains a copy of the fragment of a Web site that is visited by a user. This gives us a unique possibility to measure the CFs - the cacheability factor a Website. It is defined as follows:

$$CFs(w,R) = \sum_{x \in X(w)} \frac{CF(x,R) * hit(x) * size(x)}{size(w)}$$

where:
X(w) is the set of all pages in the TIF from a Web site w, and

$$size(w) = \sum_{x \in X(w)} hit(x) * size(x)$$

is the number of bytes of the Web site w, its popularity being accounted for.

The values of CFs for some popular polish Web sites are in the Table 6. The values of the factor differ significantly. This is due to the different nature of the Web sites but also reflects a varying degree of "cache awareness" of different Web masters.

Table 6. The CF values of the refresh rate for the selected Web sites

WebSite	RMin	R10min	RHour	RDay	Rweek	RMonth
gazeta.pl	0.74	0.73	0.69	0.44	0.25	0.08
rzeczpospolita.pl	0.83	0.78	0.73	0.63	0.49	0.35
chip.pl	0.63	0.63	0.62	0.54	0.36	0.17
wp.pl	0.60	0.58	0.54	0.34	0.18	0.07
wroclaw.pl	0.77	0.77	0.77	0.71	0.51	0.07

6. Web Content Analysis

As stated previously the browser cache contains not list of used WWW objects but also their content. This is one of the main advantages of using this sort of input data. The extraction of text from cache entries is not straightforward for all pages. The specification of the HTML format is not closely followed by many webmasters. As a result the official SGML parser of the W3C consortium available on [18] could produce as many as 300 errors or warnings while analyzing a popular Web page. The reported errors are in some cases significant e.g. the occurrence of tags that are closed but not opened.

The texts were selected from the cache files using a PERL script. The script extracted the plain text of a page divided into segments by links. Each link contained the link texts (in any) and the requested URL. In the experiment all text extracting, processing and evaluation functions were written in Perl. The moderate size of files (the maximal size of a files with selected text has not exceeded 8 MB) make it possible to process data without the use of databases.

6.1. Language Identification

It turned out that the texts were almost exclusively in Polish or English. The statistical data must be calculated for each language separately. The identification was done for each segment individually. In the process the words in a text segment were analyzed one by one. A word was classified as an English word if it was on the list of Special English Words. The list was generated in two stages. In the first stage the 2500 most frequent English word forms were selected. During the second stage the word forms that were common to both Polish and English (e.g. problem) or have by coincidence the same spelling such as "to" were eliminated. A special attention was paid to the "technical words" like "http" or "ftp" which could appear in any text. They ware assigned to a special group and were considered as neither English nor Polish. All words that were not English or "technical" were regarded as Polish words.

A text segment was considered to be Polish text if it contained more Polish than English words. The link texts were significantly shorter and for them no language identification took place.

No other natural languages were taken care of but it did not seem to influence the results in a noticeable way.

6.2. Stemming

Each language is more or less inflective and the stemming that is reducing a word to a token representing its broader meaning is considered to boost the efficiency of text processing. For that reason the extracted text were processed by stemming algorithms (stemmers) and the statistical properties of both versions the original and after stemming) were calculated and compared.

English is significantly less inflective than Polish and many efficient stemmers are publicly available. During the experiment two well known algorithms we used: the Poter [19] and Lovins [20] algorithms. The latter one is more complicated and was found more effective and therefore only it is the only one English stemmer discussed in what follows.

The stemmers for Polish are not as easily available. Fully-fledged stemmers are built into commercial products and are not freely available even for research purposes. A complete survey of Polish stemmers and the evaluation of their applicability in information retrieval was published recently by D. Weiss [21].

In the experiment a freely available although relatively old morphological analyzer called SAM 34 was used. SAM-95 is a hybrid stemmer that uses a dictionary of suffixes of inflected forms, collected by J. Tokarski. Potential stems are then generated and optionally verified against a regular dictionary of known words. The analyzer, developed by K. Szafran [22], could not be unfortunately easily integrated with other applications. Therefore, a dictionary was first constructed. A list off all different Polish word forms was collected and then it was analyzed by SAM-95. The first stem proposed for a given word form was stored in a dictionary. The dictionary was later used as an associative array in Perl scripts for text processing.

6.3. Zipf and Heaps Laws

Zipf 'sand Heaps' laws are the most known empirical laws that describe the relationship between the frequency ranks vs. the frequency of occurrence and the diversity and of vocabulary in text fragments of various lengths respectively. The laws are important from the theoretical and practical point of view. They provide us with insight into the nature of language. The natural languages comply with the laws although the coefficients used in the laws vary to some extent. From the practical point of view the laws are used while designing full-text databases or monitoring the process of language learning. The aim of the experiments was to discover the similarity of statistical properties of texts extracted from the buffer to the properties of unrestricted language. Such studies may also new directions in the field of prefetching that is reading in advance Web pages in order to reduce browser latency.

The Zipf law, published almost half a century ago [23], applies not only to natural language but also to such diverse phenomena as size of earthquakes or settlements, income distribution or frequency of accesses to Web pages. Applied to natural

language it states that in any large enough text the frequency ranks of word forms or lemmas are inversely proportional to the corresponding frequencies:

$$\log f_r \approx C - z \log r$$

where:

f_r: the frequency of the unit (word form or lemma) having the rank z

z: the exponent coefficient, usually close to 1.

Although the law applies to all natural languages the value of the coefficients differ, e.g. for English z = 0.97±0.06 and for Russian z=0.89±0.07 [24]. The difference is significant and could be partially explained by the fact that contrary to English, Russian is a highly inflective language. The study revealed that counting lamas (basic form forms) instead of word forms had hardly changed the results.

The results obtained for the TIF texts are listed in the Table 7.

Table 7. The values of Zipf law coefficients for different types of TIF texts

Buffer	Language	Stemming	z	C
INDY	Polish	No	0.92	9
INDY	Polish	yes	0.90	9
INDY	English	No	0.92	9
INDY	English	yes	0.93	9
STUD	Polish	No	0.90	11
STUD	Polish	yes	0.88	11
STUD	English	No	0.85	10
STUD	English	yes	0.98	11

Probably less known but also interesting is the Heaps law [25]. It describes the portion of a vocabulary, which is represented in text fragments of different lengths. The law can be formulated as follows:

$$\log n_i \approx D - h \log i$$

where n_i is the number of different units in a text fragment consisting of i word forms.

Table 8. The values of Heaps law coefficients for different types of TIF texts

Buffer	Language	Stemming	D	h
INDY	Polish	no	4	0.74
INDY	Polish	yes	6	0.68
INDY	English	no	3	0.66
INDY	English	yes	6	0.56
STUD	Polish	no	7	0.69
STUD	Polish	yes	5	0.70
STUD	English	no	4	0.66
STUD	English	yes	7	0.58

As in the case of the Zips law, the Heaps law applies in the general case in which we describe the objects a feature that takes distinct values. For example, the objects could be people, and the feature the country of origin of a person. If persons are selected randomly then Heaps' law says we will quickly have representatives from most countries (in proportion to their population) but it will become increasingly difficult to cover the entire set of countries.

In the cited above studies the value for the coefficient h ranged from h=0.79±0.05 for English to h=0.84±0.06 for Russian.

The Table 8 shows the values of the coefficient h for different text types that were extracted from the buffer.

The departure from the result for unabridged natural language is more clearly visible than in the case of the Zipf law. Once more again the stemming processes had changed the nature of the dependency especially in the case of English texts.

7. Conclusions and Future Work

The paper proposes the use of the local buffer as a useful source of user centric Web mining data. The main advantage of the local buffer data over the log files is that it keeps track of all user activities over a considerable long period of time. The identification of user needs and preferences could be done locally without infringing user anonymity.

Using the URL history data it is possible to identify the frequently used navigational pages. They are a good staring point for a prefeching algorithm and for local notification Websites monitors. The prefetching of content pages is however not feasible because of the varying nature of URL's for content pages and the large number of links on a navigational page. For that purpose the analysis of anchor texts is required.

The Cacheability factor gives a good measure of the susceptibility to caching. It should be stressed that although the data presented in the paper were collected locally they describe general properties of a Web site. The CF values for the examined sites differ to large extend. The next step involves the analysis the correlation between the CF, network distance and the fluctuations of loading times.

The CFs enables a Web master to compare the cache friendless of different Websites or even (which far more practical) the comparison of the CF values for different versions of the same Website. Obtaining the same information through direct measurements of latency is difficult to obtain. Direct measurement is time consuming and is influenced by many external factors.

The first results on the statistical properties of the text in show a noticeable departure from properties of the unrestricted language. For the future more detailed selection of texts is foreseen. A more precise language identification algorithm is necessary. More sophisticated stemming algorithm for Polish texts should replace the relatively old SAM-95 algorithm. The comparisons of the properties of the texts that accompany the used and unused links are a promising area of study. If the results confirm the existence of a significant discrepancy between the properties of the two types of link texts than such texts could be successfully used for prefetching.

References

[1] J. Srivastava, P. Desikan, V. Kumar: Web Mining: Accomplishments & Future Directions*, National Science Foundation Workshop on Next Generation Data Mining (NGDM'02)*, 2002.

[2] L. Tran, C. Moon, G. Thoma: Web Page Downloading and Classification, *The Fourteenth IEEE Symposium on Computer-Based Medical Systems*, July 2001.

[3] *Music Machines log data*: http://www.cs.washington.edu/ai/adaptive-data/

[4] *WorldCup98 log data*: http://ita.ee.lbl.gov/html/contrib/WorldCup.html

[5] *Log data*: http://www.ircache.net/Traces/

[6] http://www.web-caching.com/ cacheability.html

[7] Common Log Format: http://www.bacuslabs.com/WsvlCLF.html

[8] C. A. Cunha, A. Bestavros and M. E. Crovella: "Characteristics of WWW Client Traces", *Boston University Department of Computer Science*, Technical Report TR-95-010, April 1995.

[9] Gain Network: http://www.gainpublishing.com/

[10] D. Reed: Privacy and the Future of Behavioral Marketing, http://www.claria.com/advertise/oas_archive/privacy.html?pub=imedia_module

[11] A. Orlowski: *The Goo that knows you - desktop privacy branded 'unacceptable*.

[12] A. Sieminski: Changebility of Web Objects, *ISDA'05 -5th International Conference on Intelligent Systems Desin and Implementation*, Wrocław, 2005

[13] M. Rabinowich, O. Spatschech: *Web Caching and Replication*, Addison Wesley, USA, 2002

[14] X. Zhang: Cachability of Web Objects, *citeseer.ist.psu.edu/zhang00cachability.html*, 2000

[15] B. Brewington, G. Cybenko: How dynamic is the Web, *Computer Networks, Amsterdam,* Netherlands, 1999,vol. 33, Num. 1--6, pp. 257--276, url = "citeseer.ist.psu.edu/291794.html

[16] D. Wessels, *Web Caching*, O'Reilly and Associates, Inc, USA, 2001.

[17] A. Siemiński, The Cacheability of WWW Pages, *Multimedia and Network Information Systems*, Technical University of Wrocław, Poland, 2005.

[18] http://validator.w3.org/

[19] M. F. Porter: An algorithm for suffix stripping, *Program,* 14 no. 3, pp 130-137, July 1980.

[20] J. B. Lovins,: Development of a Stemming Algorithm. *Mechanical Translation and computation Linguistics*. 11 (1) March 1968 pp 23-31.

[21] D. Weiss: A Survey of Freely Available Polish Stemmers and Evaluation of Their Applicability in Information Retrieval, *2nd Language and Technology Conference*, Poznań, Poland, 2005, pp. 216-221.

[22] K. Szafran: Analizator morfologiczny SAM 95 opis użytkowy, TR 96-05 (226), *Instytut Informatyki Uniwersytetu Warszawskiego*, 1996.

[23] G. K. Zipf: *Human behavior and the principle of least effort*, Cambridge, MA, Addison-Wesley, 1949.

[24] A. Gelbukh, G. Sidorov: Zipf and Heaps Laws' Coefficients Depend on Language, *Proc. CICLing-2001, Conference on Intelligent Text Processing and Computational Linguistics*, February 18-24, 2001, Mexico City. Lecture Notes in Computer Science N 2004, ISSN 0302-9743, ISBN 3-540-41687-0, Springer-Verlag, pp. 332-335.

[25] Y. Baeza, R. Neto: *Modern Information Retrieval*, ACM Press, 1999.

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-95

Classification Visualization across Mapping on a Sphere

Veslava OSIŃSKA[a] and Piotr BAŁA[b]
[a]*Institute of Information Science and Library Studies, Nicolas Copernicus University, Toruń, Poland*
[b]*Institute of Mathematics and Computer Science, Nicolas Copernicus University, Toruń, Poland*
e-mail: wier@phys.uni.torun.pl, bala@mat.uni.torun.pl

Abstract. Existing classification schemes are visualized as hierarchical trees. Science data visualization requires a new method in information space modelling in order to reveal relations between class nodes. This paper describes a novel visualization concept of classification scheme using subject content metrics. We have mapped the document collection of Association for Computing Machinery (ACM) digital library to a sphere surface. To overcome the incorrectness of linear measures in indexes distances we calculated similarity matrix of themes and multidimensional scaling coordinates. The results show that space distances between class nodes accurately correspond with the thematic proximities. Documents mapped into a sphere surface were located according to the classification nodes and distributed uniformly. Proposed method to visualize classification scheme is proper to reach nonlinearity in subject content visualization. This property allows us to place close by more classification nodes. Symmetry of a sphere favours a new subclasses and sublevels of classification trees uniform visualization. This method may be useful in the visual analysis of Computer Science and Engineering domain development being grown instantly.

Keywords. Infovis, information visualization, science mapping, semantic browsing, automatic classification, bibliomining

Introduction

With the exponential growth of Internet resources, it has become more difficult to find relevant information from the one hand and organize information services from the other. The emphasis must to be placed on the information visualization effectiveness. The aim of these actions is to increase the perception space and cognitive capacity of user as well as to make possible an interaction between the user and an application. Artificial Intelligence research and newest discoveries in cognitive science support intelligent visualisation projects [7]. The two most powerful information processing tools: the human mind and the modern computer are connected during visualization cycle according vivid definition: "this is a process in which data, information and knowledge are transformed into a visual form exploiting people's natural strengths in rapid visual pattern recognition" [13]. This temporal information processing seems to be more related to the general intelligence level than classical measures of mental speed [5].

Current Infovis (widely used abbreviation of terms: Information Visualization) investigations relates to such fields as scientific data visualising, human-computer interaction, data analysis, data mining, computer vision and computer graphics. From this perspective Infovis developers focus on the ergonomic information systems GUI (Graphic User Interface) design. It must assure user the data browsing and retrieval as well as an easy navigation, recognizing, exploring and filtering data. Subject search (the synonym is “cross searching”) besides keywords search plays distinct role to realize the first two tasks in digital libraries environment. This approach allows the user to assist in quickly focusing on consistent and hierarchical category information. Modern digital libraries interface problems like cross search or cross browse can be engaged the classification services improving.

Evolution of some science disciplines overtakes the development of their taxonomy, especially in a newest subdisciplines and subdivisions case. This situation is noticeable in a Computer Science (CS) domain being grown instantly. A new science and technology branches with narrower or wide specialization range have appeared. However existing CS subject classification schemes do not fulfil ontological, statistic and perceptive requirements of information society [15].

Hierarchical information structures – we consider classification systems – are often visualised by means of hierarchical dendrograms. A dendrogram (from greek *dendron* "tree", *-gramma* "drawing") is a tree diagram used to illustrate the arrangement of the objects produced by a chosen criterion. Tree objects are linked by "parent-child" relations and some of them can be split. Hierarchical information is the biggest data group. The hierarchy exists in library classification systems, genotype systems, genealogy data as well as computer's directories structures and object oriented programming languages class definitions.

With a rapid expansion of information and communication technologies since previous decade the new forms of hierarchies visualising are created and evolved. The growing information space follows for the Google development and inversely. It contributed to further improvement of information visualisation methods.

1. Overview of Relative Work

From the beginning of 90th the computer processors speed fall behind with a fast growth of hard discs resources. To solve this problem the new forms of directories trees visualisation (on Unix systems) were investigated. Hierarchy trees were presented by maps instead of graphs – one dimension topology is extended to two dimensions. Files and directories multilevel structures and were illustrated in the forms of nested rectangles [17] or concentric rings [18] - TreeMap[1] and SunBurst[2] software. Another idea to magnify an exploration space in that decade was workspace construction in hyperbolic 3D geometry. The first applications which used fisheye technique (called also "focus+context") are hyperbolic browsers [10] – this approach assures more place to visualise the hierarchy. This leads to the convenient property that the circumference of a circle grows exponentially with its radius which means that exponentially more space is available with increasing distance. It is possible to study hyperbolic view of

[1] http://www.cs.umd.edu/hcil/treemap/

[2] http://www.gvu.gatech.edu/ii/sunburst/

various types of data and own samples due to on-line accessible applications like Inxight StarTree[3], Hyperbolic 3D[4] or Walrus[5].

Next decade Internet users experienced information awash which become a common technical, social and psychological problem. Inhomogeneous character of the most web resources accelerates the structural modelling research which is the second, besides graphical representation, fundamental aspect of Infovis [3], [15]. The main task of these systems is to discover complex data structure and to examine it visually. The standard way in large data set analysis is to cluster them into clusters (sub-sets) according to some similarity parameters.

Clustering techniques and classification of a large documents collection fall into two basic groups: statistical and knowledge-based techniques [12]. The first one uses the statistical linguistic, machine learning or neural network algorithms to create topical clusters. Well known, from two decades, statistical technique is self-organizing map (SOM), firstly described by Teuvo Kohonen [9]. SOM is a type of artificial neural network, it is generally applied for the objects topological classification so the spectrum of use ranges from the data mining to data visualization. The authors [16] describe self-organizing maps in the hyperbolic space and called them HSOM. These innovative simulations prove that neighbourhood neurons with hyperbolic characteristic facilitate space SOM construction.

For visualizing low-dimensional views of high-dimensional data can be used less time-consuming comparable with SOM [6] statistical method called Multidimensional Scaling (MDS). This is a data analysis method which is widely used in marketing and psychometrics. MDS rearrange objects in an efficient manner, so as to arrive at a configuration that best approximates the observed distances. It actually moves objects around in the space defined by the requested number of dimensions, and checks how well the distances between objects can be reproduced by the new configuration.

For effectively text categorization the algorithms such as Latent Semantic Analysis[6] (LSA) and Support Vector Machine (SVM) could be adopted. LSA is based on Vector Space Modeling which defines multidimensional information in the vector space. It supposes that documents collection consists of known number of clusters with hidden properties. This method searches relationships between a set of documents and the terms they contain [1], [2], [3]. The next algorithm, Support Vector Machine, merges the fields of machine learning and statistics. It maximizes the confidence margin between two classes and this way minimizes the empirical classification error. A special property of SVM[7] is the hyperplane [11] that separates a set of positive examples from a set of negative examples.

Second clustering category: knowledge based techniques relies on an explicit knowledge base such as a rule base, semantic networks, patterns, and so on. These all approaches require an extensive manual knowledge engineering effort to create knowledge base. Nevertheless these difficulties, good results in limited domain have been achieved.

Modern semantic web browsers as KartOO, TheBrain, Grokker, AquaBrower, ThinkMap developed by both corporations and academic institutions use mentioned visualization technologies. Semantic information retrieval and browning software

[3] http://www.inxight.com/products/sdks/st/
[4] http://graphics.stanford.edu/papers/h3cga/
[5] www.caida.org
[6] http://lsa.colorado.edu/, software pckages: http://iv.slis.indiana.edu/sw/
[7] http://www.support-vector-machines.org/SVM_soft.html

interfaces are designed taking advantage of the last research upon human perception. Due to interdisciplinary character of information visualization different science specialists have been involved to its studies. Joint efforts of Information Science, Computer Science, Neuroscience and Cognitive Science, Computer Graphics, Psychology and Philosophy specialists help to develop Infovis methods and technology.

2. Research Methods

Automatic classifying of documents is characterized of top-down strategy which starts from categories (classes) and then assigns items to a given categories. It is opposed to clustering which is bottom-top process. For visualization of the classification tree in 3D space the root node is to be located in the centre and all sub-nodes will spread out in all directions around of the central nodes (Walrus graphics).

Our work was focused to visualize classified documents and next to construct a new graphical representation of original classification scheme. As a target space we have chosen the sphere surface because of their symmetric and curved surface map the distances between the data more precisely than a plane. Moreover sphere is easy for navigation and retrieval processes.

Test data was derived from the digital library and all of them were classified into classes and subclasses. Data population was the highest on the lowest levels for the most classes (see 4.1 Dataset). If some sublevels nodes split conceptually the documents appeared in both nodes. We assumed that the topic similarity between classes is proportional to the number of recurrent documents. As closer thematically two subclasses the more common articles they include. This pair of classes must cross in typical dendrogram tree. And inversely dissimilar subclasses contain no common data. By counting and normalizing the number of common documents for every pair of classes and subclasses it is possible to construct matrix similarity.

Dimension of square matrix is equal to the number of all classes and subclasses occurred in the data collection. Next we used MDS algorithm to decrease matrix dimension to three. To build an optimal representation, the MDS algorithm minimized a criterion called Stress. The closer the stress approaches zero, the better is the representation.

On the basis of given coordinates we designed a classification sphere. This graphical 3D representation allowed us to visualize such attributes of class nodes as index, population, level, crossing by means of colour, size, position and transparency degree respectively. We obtained a multidimensional navigation space in which a lot of information can be conveyed in a compact display, including topics, relationships among topics, frequency of occurrence, importance, etc.

Therefore we have combined two types of techniques: statistical proximity – MDS and thematic measure of distances.

3. Experiment

3.1. Dataset

For research purposes we used a collection of publication abstracts from Classification of Computing System (CCS) digital library. CCS System was created by Association for Computing Machinery in 1964 (the next versions are published in 1991 and 1998) and developed on-the-fly. Digital ACM library includes impressive collection of abstracts and full text publications (1.4 millions text pages), ACM journals and conferences proceedings. ACM is a major force in advancing the skills of information technology professionals and students worldwide. The full classification scheme involves three concepts: the four-level tree (containing three coded levels and a fourth uncoded level), General Terms, and implicit subject descriptors[8]. The upper level consists of 11 main classes which are listed in Figure 1. Every category name start with a suitable capital: from A to K. CCS is still updating and therefore a new subdivisions are appeared with "New" label or some of existing categories are prepared to removing with suitable label "Revised". Choice of CCS literature collection was caused by two reasons. The authors are well acquainted with Computer Science domain both in practice and theoretical problems. In addition on-line access to all classified abstracts of publications has been supplied.

A. General Literature
B. Hardware
C. Computer Systems Organization
D. Software
E. Data
F. Theory of Computation
G. Mathematics of Computing
H. Information Systems
I. Computing Methodologies
J. Computer Applications
K. Computing Milieux

Figure 1. ACM CCS main classes

Every publication besides main classification may be ascribed to additional classes. CCS editors instruct authors in details how to classify their documents[9]. They have to describe the document's categories, keywords and implicit subject descriptors. Browsing document abstract they can automatically generate the tree of main and additional classifications. Multiple subclasses indicate about the wide topic content. Authors together with editors contribute to form the paradigmatic topology of documents set.

[8] http://www.acm.org/class/1998/
`[9] http://www.acm.org/class/how_to_use.html

3.2. Tools and Implementation

Database of publication abstracts was collected by a PHP 5 application – scanner running on Apache Web server. The program scans websites content to explore information about document properties in following order:

- primary and additional classification indexes
- keywords
- global terms
- data of publication
- URL address.

In the next stage data duplicates were excluded and statistical analysis was performed. Similarity matrix construction was based on unique records. For the generated distances between the nodes we have successfully applied MDS algorithm to receive MDS coordinates in 3D space. Further data processing and final visualization was realized in Matlab environment. For convenient calculation we represented data as a matrix and then Matlab was the best tool for this type of objects.

3.3. Results

Because of immensity of general ACM digital library dataset we initially investigated collection of documents published in 2007. We plan to extend range of years in the next research steps. After we rejected duplicates (the same documents appear in different classes and different levels) the total number of documents N have value 37 543. The goal of current work was focused to set this quantity of documents on a sphere surface most efficiently.

We received the final number of classification indexes including all levels and two sublevels equal 353. Next we constructed the square matrix S which consists of 353 rows and 353 columns. Suppose the rows represent the main classes and columns the additional classes respectively. Then, the element of a matrix S that lies in the i-th row and the j-th column is a number of documents belonging to both classes at the same time and written as s_{ij}: The sum of these values is equal to total quantity of unique documents in dataset:

$$N = \sum_{i=1}^{353} \sum_{j=1}^{353} s_{ij} \tag{1}$$

Diagonal elements of matrix s_{ij} i.e. $i=j$ include the number of all documents classified to suitable class. After normalization to that value we received similarity matrix with ones along the diagonal and all other elements less than one. The higher any matrix element s_{ij} the classes are more similar thematically. The biggest value of similarity 1 means the class is identical with itself.

After the similarity matrix has been converted to dissimilarity matrix (have zeros along the diagonal and non-negative elements everywhere else) 3D Euclidean space coordinates for each of 353 nodes are calculated using MDS algorithm. Stress coefficient is a measure of goodness-of-fit: how well (or poorly) a particular configuration reproduces the observed distance matrix. It has been calculated as:

$$Phi = \Sigma[d_{ij} - f(\delta_{ij})]^2 \quad (2)$$

where:

d_{ij} stands for the reproduced distances;

δ_{ij} stands for the input data;

i=1,…,353; j=1,…,353.

Obtained Stress values for 3D and 2D configurations were 0.25 and 0.3 respectively. Therefore space representation is better than plane layout for out dataset.

Objects are scattered regarding to the origin point (x=0 y=0 z=0). This position defines the coordinates of centroid for MDS dataset. It is the average of points coordinates:

$$C_{ijk} = [\Sigma x_i \, \Sigma y_j \, \Sigma z_k] \quad (3)$$

where i,j,k = 1,…353.

Figure 2 shows the distribution of objects square radius: $r_{ijk}^2 = x_i^2 + y_j^2 + z_k^2$. It can be estimated square distances gathered around the value have been obtained by least squares approximation R^2=0.47. Standard deviation - a measure of the spread of multiset's values - of radius r_{ijk}^2 is equal 0.06. How we can make certain the classes nodes are disposed in the range $\pm 0.06^{1/2}$ regarding to the surface of a sphere with R radius.

Histogram r_{ijk}^2 distribution for intervals 0.02 is illustrated on the next Figure 3. We see that majority of square radiuses is spread around the counted value 0.47.

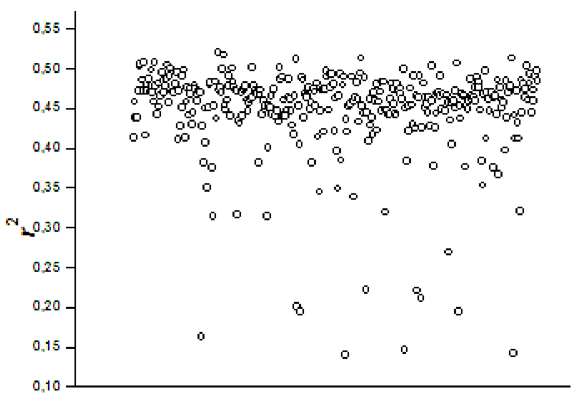


Figure 2. The distribution of square radius for given MDS data

To minimize errors in MDS pattern we erased six lowest level classification nodes consist of single document results. Therefore the objects number for visualization process decreased to 347.

Finally, all the nodes must be located on a sphere surface with radius $R = 0.47^{1/2}$. We assumed data objects can be treated like non interacting particles which have been

acting upon repelling force by the centroid node. For this goal we used Morse potential – a convenient model used in spectroscopic applications. It is a more precise approximation of the molecule's vibrational structure than quantum harmonic oscillator because it includes the effects of bond breaking, such as the existence of unbound states. Some quantum chemistry projects for atomic simulations apply Morse potential to construct energy surface - multi-timescale modelling [8].

Morse potential will have an energy minimum for R value. Near the minimum of the curve the potential energy is a parabolic function so experimental data coincide in calculated assign point.

The Morse potential energy function is of the form for diatomic molecule:

$$E_s(r) = D_e\left[\left(1 - e^{-b(r-R)}\right)^2 - 1\right] \tag{4}$$

where coefficient b and a well depth D_e were founded empirically. Required constants are obtained by making convolution of a function (4) with standard deviation of dataset.

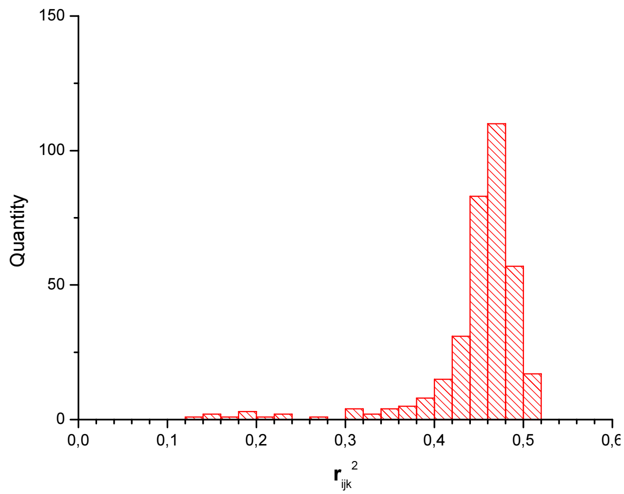


Figure 3. Histogram of rijk2 distribution for intervals 0.02

Data processing allows us to place data objects on a sphere surface representing CCS classification space. Our research approach makes the visualization of new object attributes possible. The node patch scaling on a sphere surface is accomplished according to the number of document classes. The distribution histogram for class population is shown in Figure 4.

The colour palette gives a big scale of objects characteristics. Every of 11 main classes is marked by different colour. The sublevels of classification are described by colour lightness degrees. So colour palette has been extended to the number of colours be attributed to three levels nodes: 11×3 =33. In the Figure 5 we can see the visualization of 347 classification nodes mapped to the classification surface.

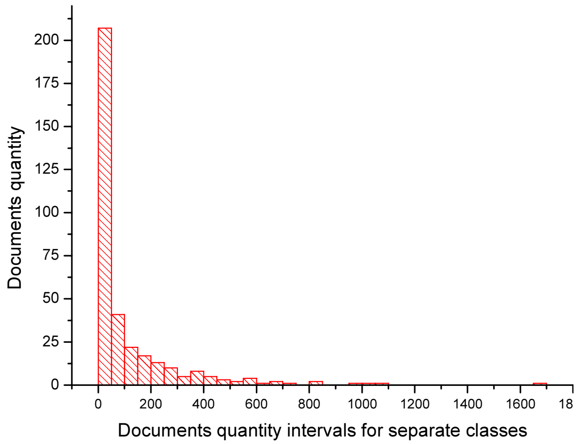


Figure 4. The distribution histogram for class population

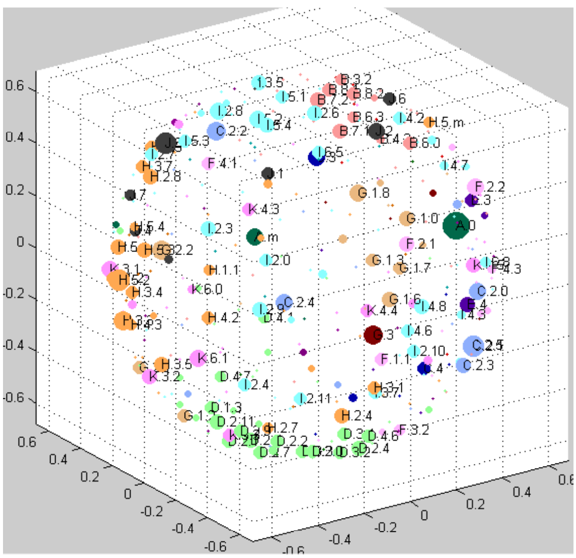
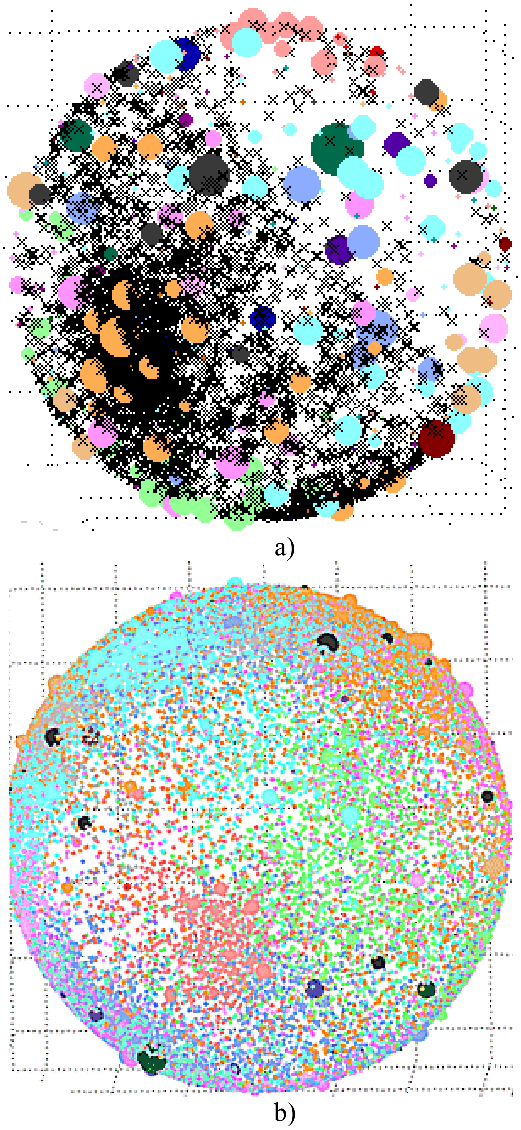


Figure 5. View of spherical classification surface

The symbols of documents must be positioned on a sphere. Every document is characterized by the main class index and various indexes of additional classes. We suppose that the weight of main classification relate to additional as 6:4. As a result every document was characterized by two attributes: the coordinate and a colour of suitable subclass. As we will see this relation influences significantly the dataset visualization. Figure 6a) demonstrates the data collection location for the class I. In this case documents marked as black symbols "+". As an evident example we can describe two contrary categories like Hardware and Software which are represented by the clusters located far each other on the two poles of sphere. As larger the same colour

clusters the proper class is more precisely described. Violet patches corresponded with Milieux class are dissipated on the whole surface. Methodologies class I characterized by a wide subject content and its clusters are occupied the most sphere space.



a)

b)

Figure 6. Documents set a) for class H (orange patches); b) all classes map

Finally all collected documents were mapped on a sphere surface observing their features. We used various text symbols and colour for class recognition. Documents inherit the colour of the lowest class as the most populated nodes. This way all 37 543 points set into a mapping 3D surface. We get a sphere shaded by 11 colour samples. As a closer inspection shows, colour patches created by colour dots were:

- of irregular form,
- scattered in different sphere places so some classes were represented by more than one colour clusters,
- with the dissolved border between colour clusters,
- arranged uniformly on a whole sphere.

Summarizing the results we note that the colour clusters correctly represent classes and subclasses properties. Articles populate 3D surface in uniform way, no empty spots were observed.

4. Conclusion and Further Research

In this work we have described the methodology we have proposed to visualize CCS classification scheme on a sphere surface. We have applied both subject content metrics and MDS technique to ACM digital library collection of documents published in 2007. Dataset counted 37543 items and the number of given classes nodes came to 353. Colour clustering on a sphere surface indicates that space distances between classes nodes accurately correspond with the topic proximities. Proximity at this point means similarity in both paradigmatic and intuitive comprehension of themes. Uniform distribution of all documents on a mapping surface ensured that this is proper strategy of examined classification tree visualization and evaluation. When a subject classification scheme evolves the classes map modelling with a preserving the distances is more possible than in classical hierarchical tree case. Nonlinear metrics allow us to place close by more classification nodes.

Moreover, the experimental environment allows us to display with zoom any surface segment of classification sphere through rotating. We can monitor topology of class nodes from any perspective. Interface provides also the representation of documents set which belongs to the chosen class. The concept of this visualization is covered with the quotation: "The eye seeks to compare similar things, to examine them from several angles, to shift perspective in order to view how the parts of a whole fit together".

In the future we intend to repeat experiment for the wider time range. Because a sphere is an easy object for data visualisation and simulation proposed model could facilitate the studies of Computer Science classification evolution. Because ACM is one of the biggest digital collections of literature in Computer Science domain the sphere allow us to observe dynamics in Information Technology and Engineering development.

Therefore it is possible to construct a classification system in a dynamic knowledge space. Space this will be useful for making historical domain analysis and prediction what subfield is far-reaching what is decayed. Classification scheme should be flexible both to visualize full coverage of a new categories intellectual content and to reduce thematic space old categories case.

For a scientist from different research fields this visualization will be very useful. Especially, on an interdisciplinary field the scientists can predict the branch growing dynamics. This investigation during a long times give us a chance to simulate a future structure of proper knowledge. On the changes dynamics research will be possible to

use some mathematical models from nonlinear dynamics for complex systems such as this data structures.

References

[1] K. Börner, et al., LVIS-digital Library Visualizer. In *CiteSeer*. Scientific Literature Digital Library. 2000. http://citeseer.ist.psu.edu/559314.html

[2] K. Börner, Visual Interfaces for Semantic Information Retrieval and Browsing. In *CiteSeer. Scientific literature Digital Library*. 2000. http://citeseer.ist.psu.edu/571532.html

[3] C. Chen, *Information visualization: beyond the horizon*. 2nd ed., Springer, London, 2006.

[4] H. Chen and S.T. Dumais, Bringing order to the Web: automatically categorizing search results. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, *Hague, Netherlands*: 145-152, 2000.

[5] J. Dreszer, et al., General intelligence and temporal control of motor tasks. *Acta Neurobiologiae Experimentalis* 67: 322, 2007.

[6] W. Duch and A. Naid, Multidimensional Scaling and Kohonen's self-organizing maps. In *Proceedings of the second conference on Neural Networks and their applications*: 138-143, 1996.

[7] W. Duch, Classification, Association and Pattern Completion Using Neural Similarity Based Methods. In *International journal of applied mathematics and Computer Science.* 10 2000, 747-766.

[8] D.D. Johnson, et al., *Multi-Timescale Modeling and Quantum Chemistry using Machine-Learning Methods via Genetic Programs*. MCC Internal Review. University of Illinois, 2005.

[9] T. Kohonen, Self Organization of a Massive Document Collection. *IEEE Transactions on Neural Networks.* 11, 2005, 74-585.

[10] J. Lamping and Raman Rao, Laying out visualizing large trees using a hyperbolic Space. *In Proceedings of the ACM Symposium on User Interface Software and Technology*, 1994, 13-14.

[11] T.-Y. Liu, et al., Support Vector Machines Classification with a very Large scale Taxonomy. In *ACM SIGKDD Explorations* 7: 2005, 36-43.

[12] K. Maly, et al., An automated classification system and associated digital library services. In *Proceedings of the 1st International Workshop on New Developments in Digital Libraries*: 2001, 113-126.

[13] I. Niskanen, *An interactive ontology visualization approach for the domain of networked home environments*.
VTT Technical Research Centre of Finland, 2007. http://www.vtt.fi/inf/pdf/publications/2007/P649.pdf

[14] J.D. Novak, et al., *The Theory underlying concept maps and how to construct hem*. Technical report IHMC CmapTools 2006-01, Florida Institute for Human and Machine Cognition, 2006. http://cmap.ihmc.us/Publications/ResearchPapers/TheoryUnderlyingConceptMaps.pdf.

[15] V. Osinska, *Semantic approach of information visualization in the net and digital libraries***.** (in polish). 2007. In EBIB. 77

[16] J. Ontrup and H. Ritter, Large-scale data exploration with the hierarchically growing hyperbolic SOM. In *Neural Networks* 19, 2001, 751-761.

[17] B. Shneiderman, *Treemaps for space constrained visualization of hierarchies*. Report. 2006. http://www.cs.umd.edu/hcil/

[18] J.T. Stasko, et al., An evaluation of space-filling information visualizations for depicting hierarchical structures. *International journal of Human-Computer Studies* 53, 2000, 663-694.

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-107

Pre-processing Techniques for the QSAR Problem

L. DUMITRIU, M-V. CRACIUN, A. COCU, and C. SEGAL
Dept. of Computer Sci&Eng, University Dunărea de Jos of Galați, Romania

Abstract. Predictive Toxicology (PT) attempts to describe the relationships between the chemical structure of chemical compounds and biological and toxicological processes. The most important issue related to real-world PT problems is the huge number of the chemical descriptors. A secondary issue is the quality of the data since irrelevant, redundant, noisy, and unreliable data have a negative impact on the prediction results. The pre-processing step of Data Mining deals with complexity reduction as well as data quality improvement through feature selection, data cleaning, and noise reduction. In this paper, we present some of the issues that can be taken into account for preparing data before the actual knowledge discovery is performed.

Keywords. predictive toxicology, Data Mining, pre-processing

Introduction

Quantitative Structure-Activity Relationship (QSAR) has been introduced in the 1960s, by investigating the relationship between the structure and the activity of chemical compounds (SAR) to support understanding the activity of interest and to allow the prediction of the activity of new compounds based on knowledge of the chemical structure alone.

Among data mining techniques the most used ones are based on neural networks [17], on neuro-fuzzy approaches [13] or on genetic programming [11]. All these approaches predict the activity of a chemical compound, without being able to explain the predicted value. Descriptive data mining techniques applied to chemical compound classification [2] may also be used for predicting toxicity classes.

The quality of the QSAR data relies on multiple readings for a given observation, for which the variation of data on the same compound should be much smaller than the variation over the series. A thorough analysis of the data may also be used to increase the quality of existing toxicological data.

1. Pre-processing techniques

Pre-processing the target dataset is a very important stage in the KDD process. The objective is to transform the initial dataset into a consistent, relevant, but manageable dataset. During this stage there are many techniques and data transformations that can be applied, but there is no recipe, since this process is highly dependent on the universe

of the data and the problem to be solved. It is also time-consuming. There are several kinds of operations:

- sampling – vertically reducing the initial dataset– is recommended whenever there is a huge amount of instances that is unmanageable when building the data models;
- noise elimination, is recommended when the initial dataset is obtained in disturbing circumstances that influence the collected data values;
- data cleaning, checks the initial dataset for inconsistent or missing values and attempts replacing those values with estimated ones;
- data integration, when several, heterogeneous data sources are used when data is collected;
- feature selection – horizontally reducing the dataset – is recommended when instances are described through too many attributes;
- discretization – reducing the size of attribute's domain for computation reasons.

For our study we have just used three of these activities, namely data cleaning, feature selection and discretization.

1.1. Data cleaning

Often, there are missing values in the dataset to be explored. There are three reasons for these values not to exist:

– lost values, meaning that the value has been known at some time in the past but it has not been preserved;
– the data source did not provide those values either because they were considered as irrelevant or obvious;
– the values could not be determined, measured or observed.

Dealing with missing values can be performed either during the pre-processing stage, thus providing the mining step with a full dataset, or during the mining stage by choosing an algorithm that can take into account missing values.

The first approach can consist of the following [5]:

- filling in missing value with the most common/probable value in the attribute's domain, like the average value or the most frequent one;
- filling in the missing value with all possible values, thus creating new instances that are added to the initial dataset;
- eliminating the instances with missing values;
- supplementing the attribute domain with a new symbol/value that has the meaning of missing value.

The second approach is materialized in algorithms like C4.5 [14] that build decision trees, where missing value datasets can be handle as they are.

1.2. Feature selection

Feature selection attempts to rank attributes with regard to their contribution to the model to be discovered.

There are three major feature selection approaches:

– filters – the selection process is independent of the knowledge discovery process,

- wrappers – the selection process is interdependent of the knowledge discovery process and
- embedded approaches – the selection process is included in the knowledge discovery process.

Since we are more interested by the pre-processing phase, we will emphasize filters.

Filters use measures of attribute quality assessment. There are context-independent and context-dependent measures [10]. The context-independent measures are fast but they completely ignore the other attributes. Context dependent measures have to trade result accuracy with performance. Some measures that offer a good compromise are the one used by the Relief algorithm family: Relief [8], ReliefF [9], RReliefF [16]. Considering that similarity measures the distance between two objects, determining the relevant feature subset consist of analyzing the lack of similarity between an attribute (or a subset of attributes) and the target attribute.

Also, Bayesian Networks [12] evaluate the importance of one attribute by observing how much the predictive performance of the model drops if we remove the corresponding variable. An important feature makes the prediction accuracy of the model to drop down when it is left out. On the other hand, if removing a feature does not affect significantly the performance of the selected model, then it is less important. The predictive performance of the network is estimated with leave-one-out (LOO) cross validation method [3, 4].

1.3. Discretization

This technique is generally applied when the value set of an attribute is too large to be manageable. As a consequence, the initial value set is mapped into a smaller value set via various transformation.

There are a large number of transformations that are applicable in order to reduce the cardinality of the initial value set. One can use an interval membership function, equally or not distributed over the value range, or a clustering procedure, or a value frequency analysis etc.

2. Real-World QSAR problem

The initial dataset used for the study of QSAR for predicting the toxicity of pollutants consists of 184 substances. These substances are characterized, with data provided by the online database ChemID Plus (Information resources and services in toxicology, http://chem.sis.nlm.nih.gov/chemidplus/), by 266 descriptors and 182 measured toxicological effects on almost 20 species of birds and mammals.

The toxicological effects of each substance are described by the lethal dose LD50 on each species, associated with the administration manner of the toxic substance to the individuals (ingestion, inhalation, dermal contact etc.).

The descriptors are grouped in 5 categories: structural, geometrical, topological, electrostatic, and quantum-mechanical.

One issue consists of the multitude of toxicological effects due to the fact that only one model can not predict all these effects, so several models would have to be built. In this paper, we attempt building one predictive model, for one toxicological effect. In

order to choose one toxicological effect to be modelled, out of the 182 that are measured, we consider a selection procedure based on dependency analysis.

Another issue is related to large number of missing values for those toxicological effects, otherwise known as target attributes, over the substance set, so selecting those with reasonable substance coverage has to be performed.

2.1. Target attribute evaluation

We have checked the data dependencies between these toxicological values using Bayesian networks. Bayesian networks allow computing the probabilities for the dependency relationship between variables. High dependency probabilities lead to the conclusion that some of the data are redundant.

The assessment of attribute quality, as well as finding dependencies within the data is performed with B-Course (Myllymäki et al., 2002). B-Course (b-course.hiit.fi), a web-based tool for causal and Bayesian data analysis is provided by the „CoSCo – Complex Systems Computationgroup of „Helsinki Institute for Information Technology.

First stage of dependency computation has been applied only to toxicological effects data and ruled out 55 toxicological effects because they showed independence with all others, meaning that they can not enrich the model. The interpretation of this decision considers that a substance is deemed toxic if it shows a toxic effect in general. A defective measurement process may be the cause of highly different toxicological effects of the same substance when administered in various manners to various species. Since we have no information on the way the toxicological data have been obtained, we rule out the data that doesn’t show a level of consistency for the same substance.

On the remaining 127, a second stage of computation has been performed. Other 84 toxicological effects have been eliminated because they show strong dependencies with the rest, meaning that they are redundant. At this stage, the model allows us to select a representative toxicological effect out of a class of effects with similar behaviour.

The remaining 43 effects have been re-analyzed and other 11 have been eliminated as redundant.

Due to the lack of coverage of the substance set, only 6 toxic effects have been kept for the data pre-processing stage, namely: mouse - oral and intra-peritoneal, rabbit – dermal contact, and rat- oral, intra-peritoneal and dermal contact. The other toxic effects had too few measured values for the substance set of interest for the problem.

2.2. Data cleaning

In our dataset we had 591 toxicological values, out of the 1104 needed (6 lethal dose values, corresponding to the 6 selected toxicological effects, for 184 chemicals), meaning 53.5% coverage.

To fill in the missing value for a toxicological effect T corresponding to a substance S, we have used a similitude-based method. We have considered identifying the neighbours of S, as the ones for which in at least 50% of the cases, the value for the toxicological effects that have measured values, others than T, varies in average with at most 5%. We used an approach based on a partial distance function between instances in order to compute the similitude degree with S; the missing value attribute for the S

being filled in with the value of T from the nearest neighbour of S. Using this strategy, we could estimate other 162 values, thus the coverage percentage reaching 68%.

This dataset has been split in 6 different datasets, one for each of the toxicological effect as target attribute.

These 6 datasets will be used to assess the predictive capability of the 266 descriptors, in order to rank them from the relevance point of view.

The 6 datasets comprise as follows:

a) mouse LD50 intra-peritoneal: 55 chemicals;.
b) mouse LD50 oral: 68 chemicals;
c) rabbit LD50 dermal contact:56 chemicals;
d) rat LD50 intra-peritoneal 52 chemicals;
e) rat LD50 dermal contact: 52 chemicals;
f) rat LD50 oral: 76 chemicals.

2.3. Feature selection

We use an extension of the partial distance function used in the Relief algorithms to evaluate the predictive capability of numeric or symbolic attributes [10] as well as the ones with linguistic values [1]. In the end, each attribute is ranked according to its predictive capability.

The partial distance function is set to a neighbour radius of 0.1 and an indiscernability threshold of 0.01. Other methods (provided by Weka, from Waikato University) will be used for benchmarking: BestFirst -a filter method, Wrapper and CART-as an embedded method.

Also, all these methods are 10-fold cross-validated and the final selected features for each dataset is resulting from a vote between the results from each selector.

The results obtained in what concerns the feature selection show an interesting correlation between the relevant attributes associated with the same administration manner of the substance, regardless of the specie they correspond to. They also reduce the space of 266 descriptors to sets from 30 to 50 relevant descriptors.

2.4. Target attribute discretization

We are interested in the discretization of the target attribute, namely the toxicological activity of a compound.

For each chemical, the value of the lethal dose has been initially discretized by conversion into a toxicity degree according to the Hodge & Sterner scale (see Table 1).

Table 1. Toxicity degrees on Hodge&Sterner scale

Degree	Label	LD50 (mg/kg)
1	High	< 1
2	Regular	< 50
3	Moderate	< 500
4	Low	< 5000
5	Very low	< 15000
6	Harmless	> 15000

2.5. Pre-processing result validation

In order to validate the results, the 6 datasets and the associated relevant attribute sets have been presented to chemistry specialists. At this moment, the datasets are more readable, due to the relevant attribute sets.

The chemists have selected as target attribute toxicological values for rat LD50 oral administration. They have also decided, according to the ranking of the predictive capability of the attributes obtained from feature selection, to 4 different sets of descriptors for the 66 chemicals: one with 52 descriptors, one with 34, one with 26 and one with 13 descriptors. These sets are overlapping.

Due to the complexity of physiological processes that can determine a toxic effect, the training set has been populated, by the chemists, with 45 of the chemicals and the test set with the rest of 21. The 45 chemicals from the training set are selected in order to cover different intoxication mechanisms within the body.

3. Experimental Validation

With the training and test sets, as well as the discretized target attribute, we attempted toxicological degree prediction through classification.

The following classification methods were used:

- Artificial neural network implemented in Matlab. The network structure is based on multilayer perceptrons with back propagation with an input layer containing the same number of neurons like the number of descriptors, a hidden layer with half of the number of inputs and outputs sum, and an output layer with a single neuron.
- Neuro-fuzzy interference system ANFIS [7] use for adjusting the parameters of membership functions a learning process similar with neural networks combined with fuzzy inference rules engine.
- Classification and regression tree algorithm CART [14] divides recursively the data using an hierarchic structure that represents decision trees. Implementation was made with implicit Matlab functions.
- Naïve Bayes classifier [15] is a simple probabilistic classifier based on applying Bayes' theorem with naive independence assumptions. The resulting probability model is an independent feature model. The model construction and the assessment of attribute quality is performed with B-Course [12]. B-Course is a web-based tool for causal and Bayesian data analysis. Also, we use Weka [18] for testing reasons, because the b-Course does not support automatic classification accuracy calculation with a test dataset.
- NBTree [6] is a hybrid decision tree with naïve Bayes classifiers in leafs. This method uses decision tree as the general structure and deploys naive Bayesian for classification purposes. In the literature is stipulated that naïve Bayesian classifiers work better than decision trees when the sample data set is small.

The best results are obtained with NBTree approach (see Table 3), but it did not score very high on the training test (see Table 2). Naïve Bayes is second, but it has the same problem in fitting the training data. ANFIS offered the worst results on the test data, due to the obvious over fitting during training.

Table 2. Prediction accuracy – training set

Methods	Set no. 1	Set no. 2	Set no. 3	Set no. 4
Neural Net	33 / 45 (73.33%)	18 / 45 (40.00%)	28 / 45 (62.22%)	23 / 45 (51.11%)
ANFIS	45 / 45 (100%)	45 / 45 (100%)	45 / 45 (100%)	45 / 45 (100%)
CART	**41 / 45 (91.11%)**	41 / 45 (91.11%)	41 / 45 (91.11%)	43 / 45 (95.56%)
Naïve Bayes	31 / 45 (68.89%)	31 / 45 (68.89%)	33 / 45 (73.33%)	27 / 45 (60%)
NBTree	30 / 45 (66.67%)	31 / 45 (68.89%)	32 / 45 (71.11%)	26 / 45 (57.77%)

Table 3. Prediction accuracy – test set

Methods	Set no. 1	Set no. 2	Set no. 3	Set no. 4
Neural Net	10 / 21 (47.62%)	11 / 21 (52.38%)	9 / 21 (42.86%)	11 / 21 (52.38%)
ANFIS	10 / 21 (47.62%)	7 / 21 (33.33%)	8 / 21 (38.10%)	5 / 21 (23.81%)
CART	**12 / 21 (57.14%)**	10 / 21 (47.62%)	10 / 21 (47.62%)	5 / 21 (23.81%)
Naïve Bayes	10 / 21 (47.62%)	14 / 21 (66.67%)	14 / 21 (66.67%)	9 / 21 (42.86%)
NBTree	15 / 21 (71.43%)	6 / 21 (28.57%)	6 / 21 (28.57%)	5 /21 (23.8%)

The accuracy favours the larger dataset excepting the Naïve Bayes approach that performs better over the 3rd set. It also shows CART as the best performer, since ANFIS is artificially high due to the over fitting in the training data.

Moreover, analyzing the confusion matrices on these classification models, it has been observed that the Hodger&Sterner scale conversion of the target attribute is not adequate to our prediction purpose, so we are considering new experiments with fewer toxicological degrees.

4. Conclusion and Future Work

The pre-processing stage we have considered is mostly due to the fact that large amounts of data are inappropriate to be submitted to biochemists for interpretation. Feature selecting, as well as filling missing values, is helpful in extracting essential

data from the initial dataset. The reduced dataset and ranked descriptors are a basis for data interpretation by specialists.

We are considering devising a method for the confusion matrix analysis, allowing a better class definition of the target attribute labelling.

Acknowledgments

This work was partially funded by the TOPAR Romanian project, under the contract CEEX/RELANSIN/535-13/09/2006.

References

[1] M. V. Crăciun, A. Cocu, L. Dumitriu, C. Segal, A Hybrid Feature Selection Algorithm for the QSAR Problem, *Lecture Notes in Computer Science*, Springer, Berlin / Heidelberg, Vol. 991, 2006, 72 – 178.

[2] M. Deshpande, M. Kuramochi, N. Wale, and G. Karypis, Frequent Substructure-Based Approaches for Classifying Chemical Compounds in *IEEE Transaction on Knowledge and Data Engineering*, Vol 17(8), 2005, pp. 1036-1050.

[3] P. Domingos, M. Pazzani, Beyond Independence: Conditions for the optimality of the simple Bayesian classifier, *Proceeding of the 13th ICML*, 1996.

[4] C. Elkan, *Naïve Bayesian Learning*, Technical Report, University of California, 1997.

[5] J. W. Grzymala-Busse, Three Approaches to Missing Attribute Values – A Rough Set Perspective, *4th Int. IEEE ICDM'04*, 2004, 57-64.

[6] L. Han, Y. Yuhong, *Learning Naive Bayes Tree for Conditional Probability Estimation*, CAI, 2006.

[7] R.J.S. Jang, , ANFIS: Adaptive network-based fuzzy inference system, *IEEE Trans., Man and Cybernetics*, 23, 1993, 665-685.

[8] K. Kira, L. A. Rendell, *A practical approach to feature selection*, ICML, Morgan Kaufmann, 1992, 249-256.

[9] I. Kononenko, Estimating attributes: Analysis and Extension of Relief, *Proc. of ECML'94*, Springer-Verlag, 1994, 171-182.

[10] I. Kononenko, Evaluating the quality of the attributes, *Advanced Course on Knowledge Technologies*, ACAI, Ljubljana, 2005.

[11] W. B. Langdon, S. J. Barrett, Genetic Programming in Data Mining for Drug Discovery, *Evolutionary Computing in Data Mining*, Springer, Ashish Ghosh and Lakhmi C. Jain, , 2004, 211--235.

[12] P. Myllymäki, T. Silander, H. Tirri, P. Uronen, B-Course: A Web-Based Tool for Bayesian and Causal Data Analysis*, International Journal on Artificial Intelligence Tools*, Vol. 11, No. 3, 2002, 369-387.

[13] C.D. Neagu, , E. Benfenati, G. Gini, P. Mazzatorta, A. Roncaglioni, Neuro-Fuzzy Knowledge Representation for Toxicity Prediction of Organic Compounds, *Proceedings of the 15th ECAI*, Frank van Harmelen (Ed.), ECAI'2002, Lyon, France, 2002, 498-502.

[14] J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kauffman, 1993.

[15] I. Rish, An empirical study of the naive Bayes classifier, *IJCAI Workshop on Empirical Methods in Artificial Intelligence*, 2001.

[16] S.M. Robnik, I. Kononenko, An adaptation of Relief for attribute estimation in regression, *Proc. of the 14th ICML*, Morgan Kaufmann, 1997, 296–304.

[17] Z. Wang, G. Durst, R. Eberhart, D. Boyd, Z. Ben-Miled, Particle Swarm Optimization and Neural Network Application for QSAR, *Proceedings of the 18th IPDPS*, Santa Fe, New Mexico, 2004.

[18] I.H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann Publishers, San Francisco, CA, 1999.

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-115

STAH-TREE: Quality Tests of Hybrid Index for Spatio-Temporal Aggregation

Marcin GORAWSKI, Michał GORAWSKI, and Michał FARUGA
Institute of Computer Science, Silesian University of Technology, Poland

Abstract. This paper presents a new method for storage and access to spatio-temporal data. That is spatial objects that have some non-spatial attributes updated asynchronously. An example of such objects may be water meters. Proposed method is a hybrid of well documented dedicated solutions for spatial, temporal, spatial aggregate and temporal aggregate data processing. Thanks to that it was possible to achieve high performance for detailed and aggregate query processing without usage of approximation. Index name (i.e. STAH-tree) is English abbreviation and can be extended as Spatio-Temporal Aggregation Hybrid Tree. Part of this work aims in creation of cost model checked against experimental results of system performance. Some other experiments that verify system behavior were also performed.

Keywords. spatio-temporal indexing, data warehousing, spatio-temporal data warehouse, distributed data warehouse

Introduction

Nowadays, one of the important challenges that data oriented system designers have to deal with is efficient retrieval of spatio-temporal aggregates. Relational systems are almost unusable for this purpose as number of data that are involved is enormous. Considering a telemetric system for middle sized town that manages 50,000 meters (for example water meters). Asynchronous measurements updates are transmitted in one hour time frame. After a month, system stores 36 million measurements. These measurements are not easy to index as their key is multidimensional (consists of meter id and timestamp when value was measured). Multidimensional index makes it highly inefficient to use hash or one dimensional index like B-tree and these methods are wide-spread in relational database systems. This problem is multiplied by need to select subset of spatial data that is also done based on multidimensional key (X and Y coordinates are used). As a result it is required to perform join on large not indexed tables in order to retrieve answer for spatio-temporal query. Such operations take long time to run and consume large amount of system resources (CPU time, memory used for temporal results and disc accesses). Moreover, when aggregate query is taken into consideration, detailed data are retrieved only to compute aggregate (for example one of typical aggregations may be sum of water consumption in specific region). Thus this approach leads to not necessary waste of system resources.

All above mentioned issues leads to conclusion that invention of solution that accelerates spatio-temporal aggregate queries processing is not only an interesting challenge from computer-science area but it is also a commercial target. Solution that

manages larger number of users, offers shorter reaction time or/and does not require very powerful hardware is leading to cost reduction and/or increases incomes.

It is an extension of [1] that presents system for efficient processing of temporal and spatial queries as well as detailed and aggregate spatio-temporal queries. Elements that are building blocks of proposed solution are widely described in literature. Spatial data processing with usage of R-tree Quad-tree families is described in [2], [4], [5] and [7]. Introduction of spatial aggregate retrieval with usage of AR-tree and aP-tree can be found in [4] and [7]. 'Multiversion' and 'overlapping' techniques that are used for temporal data processing are covered in [3], [6] and [8]. There are documents that introduce moving object processing with usage of TPR*-tree and those that cover temporal aggregates processing. But there was no paper that aims in detailed and aggregate queries over spatio-temporal data.

The proposed solution is derived from before mentioned techniques either directly (structure that is build based on R-tree and MVB-tree) or indirectly (the manner in which aggregates are stored and computed was created based on aP-tree). Most of these techniques were slightly modified to fit in the structure. The main power that stands behind system efficiency is close cooperation of all components.

The paper puts stress on showing results of experiments that were performed on STAH. Experiments are focused on: (1) test acceleration that is introduced by additional techniques introduced in [1] (AGG, MAP, BATCH); (2) examine influence of data parameters (number of spatial objects, number of temporal updates) on index performance; (3) study influence of index internal properties on system performance; (4) examine accuracy of proposed cost model. Results of these experiments are placed in paragraph 4. Paragraph 5 contains conclusions that come from experiments as well as a summary of whole material covered by article.

1. Introduction to Index Structure

STAH-tree was introduced in article [1]. Information placed there are a good (and almost required) introduction to material placed in this article. It also contains description of three main modifications of the base structure. These modifications may or may be not applied to index (as they introduce some limitations as well) and have great impact on system performance. The index structure will only be described shortly here. In order to know details, please refer to above mentioned paper.

STAH-tree is a hybrid of spatial index (R-tree) and temporal index (MVB-tree). These indexes show high performance in their area (i.e. managing spatial or temporal data).

Base version of this system (called BASE in short) is not introducing any big structural modifications. Spatial data are stored in R-tree, temporal data are stored in MVB-tree. Spatio-temporal query is first referring to R-tree in order to retrieve set of spatial object identifiers that fulfil the query. This list is used to pose queries on MVB-tree. Data in both trees are related with spatial object identifiers that are used as measurements identifiers in MVB-tree. In order to retrieve aggregate values MVB-tree has to pose two timestamp queries (first at the beginning on time frame, second at the end) and compute results based on these values.

Conference paper [1] introduces three main modifications (BATCH, MAP and AGG) that allow accelerating query processing multiple times and tighter binding

among system components. System that is using all above mentioned modifications (called FULL) has response time up to hundred times shorter that BASE version.

The most important modification is AGG – that uses modified AR-tree in place of R-tree. MVB-tree is filled not only with measurements for objects – but when such update arrives every node that contains this object in AR-tree (one per level) update its aggregate as well. Space consumption rises – but now system is able to accelerate large queries greatly. When query contains some node – then aggregate for this node is taken from MVB-tree and deeper traversal into its sub-tree is not required. Thanks to this acceleration between system with AGG and BASE system follows this between R-tree and aR-tree when spatial queries are taken into consideration.

BATCH technique is posing large impact to number of node accesses and increase CPU usage. This is because larger set of temporal queries (sorted spatial object ids) is posed in single temporal batch query. When single node contains more than one value – node does not have to be accessed multiple times. Its impact on query answer time is not too big when additional buffers are used.

MAP technique requires constant set of spatial objects. That is – spatial index has to be loaded once and this has to be done before any MVB-tree update is posed. It changes object identifiers in R-tree that nodes that all elements in node have consecutive numbers and elements that lie in sibling nodes have numbers near each other. This introduces order in MVB-queries and retrieves acceleration either with usage of buffers or BATCH technique.

2. Cost Model

Predicting cost of query execution with few simple computations and structure parameters (node capacity etc.) along with basic data properties (number of spatial objects, number of temporal updates etc.) can be very useful. Thanks to this it is possible to optimize query execution in systems that provide alternative query execution paths. Example of such alternative execution path may be direct query over relational database. It is also possible to estimate query execution time and present this value to user. This may have big impact on user option about the system (especially in case of long least operations).

Model that is proposed in [1] is restricted by following assumptions: (1) AGG modification is used; (2) spatial data have uniform distribution; (3) buffering is not taken into consideration (as a result number of node accesses is higher than it will be in real system); (4) BATCH and MAP techniques are not used; (5) aggregation values are used only on leaf level – as this is easier to present in formula – and it is posing errors only for large queries; (6) model is showing performance for spatio-temporal aggregation queries. Such model can be described as 'pessimistic model' as always worst assumptions were taken. Real system performance will always be better than this model predicts.

Model consists of two separate parts. Formulas that were derived for MVB-tree (for example in [3]) are the first component. Model is using formula for number of MVB-tree node accesses while executing timestamp query on this index (equation 1). Second component are formulas for spatial query selectivity among spatial objects (equation 2) and aR-tree nodes (leaves) (equation 3). Overall query execution cost is computed as total selectivity (sum of selectivity for spatial objects and aR-tree leaves) multiplied by cost of two queries on MVB-tree. Final version is presented in equation 4.

Part of formulas that was used in computation is taken from literature ([2], [3], [6] and [7]). Formulas are presented in simplest form. Final equations along with their derive can be found at [1].

$$NA_{\sin gle} = \lceil \log_b N \rceil \tag{1}$$

$$SEL_o = P_p \cdot N_1 \cdot a \cdot f \tag{2}$$

$$SEL_1 = P_z \cdot N_1 \tag{3}$$

$$NA_{MVB} = 2NA_{\sin gle}(SEL_0 + SEL_1) \tag{4}$$

Table 1. Symbols that are used in equations

Symbol	Description
J	Tree level. j=0 for spatial objects; j=1 for leaves level
N	Number of spatial objects
N_j	Number of objects at j level
f	'fanout' – number of objects in node
P_z	Probability that object is contained in query range
P_p	Probability that object is crossing query
a	Factor used in describing how large part of objects falls in query range when leaf is crossing with query
b	MVB-tree node capacity
NA	Number of node accesses
SEL	Selectivity

The largest disadvantage of proposed model is assumption of spatial-data uniformity. Many experiments were performed in order to derive model that supports arbitrary data distribution – but they were not successful, as they required detailed information of aR-tree structure (node locations, node sizes, etc.). The method that is based on spatial partitioning and partial results computation (proposed in [2]) fails as STAH uses aggregates on higher levels of aR-tree (and these values changes when only partition of whole space is taken into consideration). Tree structure depends not only on data distribution but also on order in which spatial objects are inserted. Correct 'model' was introduced – but it requires summary information in aR-tree nodes (node localization, number of elements in nodes, number of nodes on specific level, etc.) This 'model' is not presented here as: (1) it is not pure mathematical model; (2) implementation is just a simple extension to R-tree structure.

3. Experiments

Experiments were performed on PC computer with AMD Athlon 2500+ processor, 1GB RAM and hard drive with SATA 133 interface. Main memory size has small influence on query processing performance as LRU buffering was used. Small number of nodes was stored in memory.

As long as not directly stated in experiment description, following assumptions are held: (1) MVB-tree node capacity is set to 200; (2) aR-tree node capacity is set to 50; (3) SVO for MVB-tree is set to 150; (4) SVU for MVB-tree is set to 2; (5) number of spatial objects is 25000; (6) number of updates for each spatial object is 100; (7) spatial query size is in 1-81% range of whole spatial area (square of 1 to 90 % in each direction); (8) temporal query range is in 2 to 5 average time between two consecutive updates. Each result is an average from 1000 experiments.

3.1. AGG, MAP ant BATCH modifications impact on query performance

System was modified with usage of three techniques: AGG, BATCH and MAP. Acceleration that is introduced by each of these techniques is shown by comparing FULL system with FULL system without this technique. For example, BATCH technique performance is shown by comparing FULL system to system that uses only AGG and MAP techniques.

BATCH technique requires batch size to be matched with I/O and CPU performance of system that it is executed on. That's because there is a batch size limit where CPU cost dominates I/O cost – and acceleration degrades. For test system this limit is B=50 where B is number of elements in batch. BATCH technique decreases number of node accesses (up to B times). This is not directly transferred into time acceleration as LRU buffers were used. Results for these experiments are presented on figure 1.

Assigning spatial identifiers according to location in leaves of R-tree (i.e. MAP technique) in introducing order in queries posed to temporal index. That leads to good acceleration results especially when BATCH technique is used. Acceleration (in node accesses and execution time) is reaching 300% for most demanding queries (query area in 40-60% of whole spatial area). Results for this technique are presented on figure 2.
AGG technique is using aggregate capabilities of aR-tree. It gives great acceleration in execution time (as well as in number of node accesses). Acceleration reaches 2000% in execution time and 3000% in number of node accesses. It is most important technique that was proposed in [1] to increase system performance in spatio-temporal aggregate query processing area. Results for this experiment are presented on figure 3.

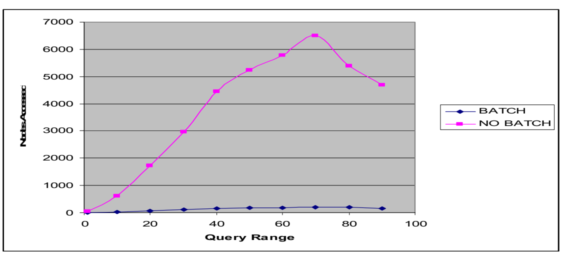


Figure 1. Number of node access in function of spatial query size for systems with and without BATCH technique

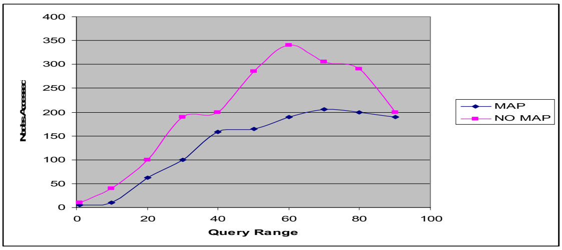


Figure 2. Number of node access in function of spatial query size for systems with and without MAP technique

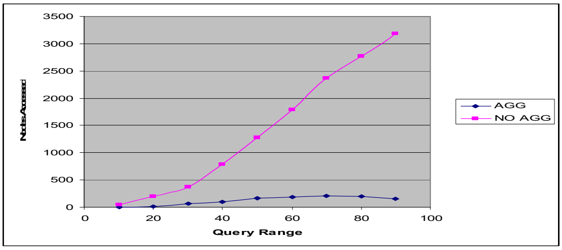


Figure 3. Number of node access in function of spatial query size for systems with and without AGG technique

3.2. Query, tree and data parameters influence on system performance

This sub-point contains results of experiments that examine influence of different parameters (tree, data and query) on overall query performance. This experiments use FULL version of STAH-tree.

3.2.1. Temporal query range

Target for this experiment is to prove that performance of spatio-temporal aggregation query is independent of range of temporal query. This conclusion is taken directly from analysis of query algorithm, as no matter how large time range is, it will always be answered with usage of two timestamp queries at the ends of temporal range. These assumptions were proved by experiments. Results are presented on figure 4.

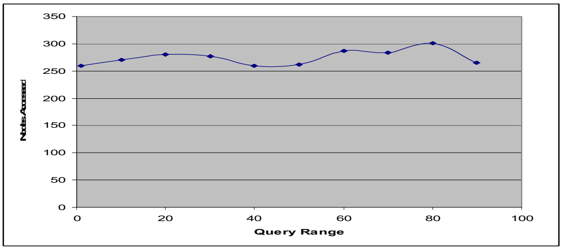


Figure 4. Number of node access in function of temporal query range

3.2.2. Number of updates per spatial object

This experiment aims in proving that system performance does not depend on number per updates for single object (i.e. history size). Number of updates has influence on MVB-tree root list size but does not change tree height. Tests were performed for 40 and 100 updates per object. Figure 5 presents gathered results. Number of node accesses is almost identical and that is the proof.

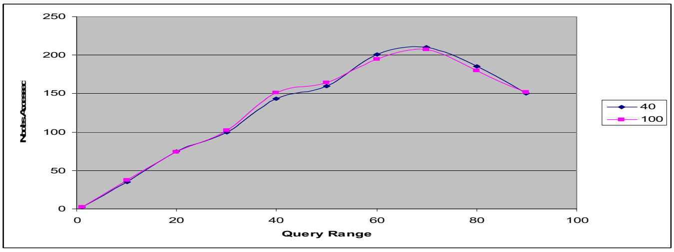


Figure 5. Number of node accesses in function of temporal query range for systems with different number of updates per object

3.2.3. Number of spatial objects

Number of spatial objects indexed by STAH-tree has double impact on index performance. First, MVB-tree height grows along with keys count growth. This growth is very small (logarithmic with base equals node capacity) as it is in case of B-tree. Second, aR-tree structure, number of nodes and spatial objects returned as query answer also grows with total number of spatial objects. Number of this object has direct impact on number of temporal queries that have to be posed. Experiment results are presented at figure 6. Highest differences can be found for 'average' queries (40%-60%). For larger queries differences are amortized with aR-tree aggregate properties (more objects give smaller nodes that easier fall into query range).

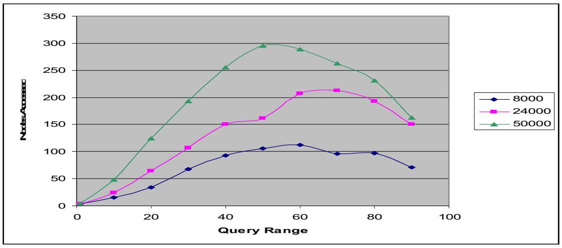


Figure 6. Number of node accesses in function of query range for systems containing different number of spatial objects

3.2.4. aR-tree node capacity

Aggregate properties of spatial index have large impact on system performance. In this experiment STAH-tree performance is checked with changing aR-tree node capacity. Capacity is maximum number of object that can be directly referenced in tree node. This parameter determines node size – and that has great impact on containment in query range probability for node. Larger capacity leads to larger nodes. Because of larger nodes, number of objects returned by spatial query is larger. Optimal solution is

to choose such capacity that maximizes node usage (fan-out as close to capacity possible) with required tree height. For example for node capacity 50 and 100 set of 50000 objects will generate trees of the same height (3) but node usage will be significantly worst for latter one. Results for this experiment are presented on figure 7.

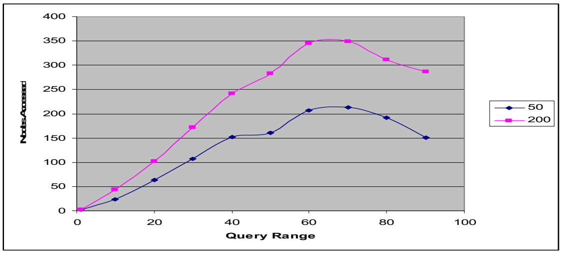


Figure 7. Number of node accesses in function of query range for systems that have different aR-tree node capacity

3.3. Cost model correctness

Experiments that verify cost model accuracy are limited to selectivity among spatial objects and aR-tree leaves. Formulas for MVB-tree were taken from literature ([3], [6]) where experiments proving their accuracy are also located. For MVB-tree tests were performed only to verify implementation correctness.

3.3.1. Selectivity among tree nodes (leaves)

This experiment consists of two separate tests. Results for both tests are presented on figure 8.

First, aims in validation of model correctness. That is, verification if experiments performed on modified aR-tree gives similar results to compute with model usage. Modified means that it allows aggregation on leaf level only (as it's assumed for model). Experiment proved almost 100% correctness.

Second, aims in validation of model accuracy. That is, verification if model follows selectivity in real aR-tree. Real, means that aggregations is performed whenever possible. Overestimation takes place for large queries (as it was predicted before). That lowers accuracy significantly. In order to increase accuracy, model shall be extended with aggregations on higher levels. This result may be treated as pessimistic value for large queries. That is, computed result will be always worse than real one. For smaller ones, model follows real values.

Inaccuracy for larger queries has not large impact on overall model accuracy as number of nodes is dominated by number of objects. And prediction for objects selectivity is precise.

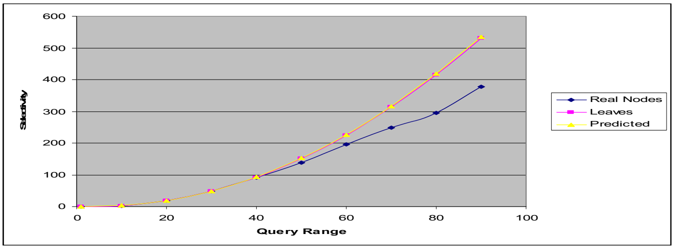


Figure 8. Correctness and accuracy of provided selectivity model. Real nodes stands for real value. Leaves stands for nodes accessed in modified aR-tree. Predicted is computed value

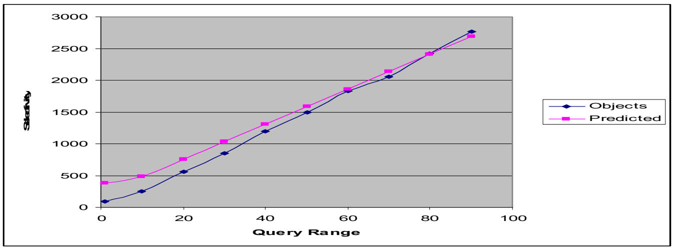


Figure 9. Accuracy of proposed selectivity model for spatial objects. 'Objects' stands for measured value. 'Predicted' stands for value that was retrieved from model

3.3.2. Selectivity among spatial objects

Model accuracy was checked against two dimensional data. For every experiment (different number of spatial objects, different node capacities and different sizes of spatial objects) results were accurate. Average error ranges from 1-5%. Largest differences are recognized at ends of query sizes (1 and 90%). Experimental results are presented on figure 9.

4. Further Work

STAH-tree is a system that manages spatio-temporal data and provides techniques for efficient spatio-temporal query processing. Spectrum of available queries is quite large it contains: 'standard' spatial queries (range query, K-nearest neighbors, etc.); aggregate spatial queries (range aggregate); temporal queries (timestamp query, time range query, time range aggregate query, etc.); detailed spatio-temporal queries (history of measurements from specific district in specific time range); spatio-temporal aggregate queries (sum of gas usage from specific district in specific time range). System performance for queries that operate on standalone indexes (R-tree, MVB-tree) is identical as if they were considered separately. Performance for spatio-temporal queries was presented in experiments. It is worth mentioning that system performance is (to some extent) independent from data parameters that have large impact on performance in 'standard' systems. For example – number of updates has no impact in STAH-tree but in relational system is very important (joins on large tables). Thanks to that system is able to assure high performance no matter how large history shall be considered. Article presents also experiments on accuracy and applicability of selectivity model for spatial objects and nodes of aR-tree. This model is accurate (as far

as uniform data distribution is considered). With those available in literature ([3], [6]), the cost model for MVB-tree creates cost model for spatio-temporal aggregate queries.

References

[1] M. Gorawski, M. Faruga, STAH-tree. Hybrid Index for Spatio Temporal Aggregation, *9th International Conference on Enterprise Information System (ICEIS2007)*, Portugal (2007).

[2] Y. Theodoridis, T. Sellis, A model for the Prediction of R-tree Performance. *Proc. Symp. Principles of Database Systems* (1996).

[3] Y. Tao, D. Papadias, J. Zhang, Efficient Cost Models for Overlapping and Multi-Version Structures. *TODS* (2002), 27(3), 299-342.

[4] Y. Manolopoulos, A. Nanopoulos, A.N. Papadopoulos, Y. Theodoridis, R-Trees, Theory and Applications. Springer (2006).

[5] N. Beckerman, H.P. Kriegel, R. Schneider, B. Seeger, The R*-tree: An efficient and robust access method for points and rectangles. *Proc. SIGMOD International Conference on Management of Data* (1990), 322-331.

[6] B. Becker, S. Gschwind, T. Ohler, B. Seeger, O. Windmayer, An Asymptotically Optimal Multiversion B-tree *VLDB Journal* (1996), 5(4), 264-275.

[7] Y. Tao, D. Papadias, Range Aggregate Processing in Spatial Databases. *IEEE Trans. Knowl. Data Eng.* (2004) 16(12), 1555-1570.

[8] J. Bercken, B. Seeger, Query Processing Techniques for Multiversion Access Methods. *VLDB* (1996).

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-125

An Attempt to Use the KEEL Tool to Evaluate Fuzzy Models for Real Estate Appraisal

Tadeusz LASOTA[a], Bogdan TRAWIŃSKI[b], and Krzysztof TRAWIŃSKI[b]
[a]*Wrocław University of Environmental and Life Sciences, Poland*
[b]*Wrocław University of Technology, Poland*
e-mail: bogdan.trawinski@pwr.wroc.pl, tadeusz.lasota@wp.pl, krzysztof.trawinski@vp.pl,

Abstract. Fuzzy models to assist with real estate appraisals are described and previous experiments on optimizing them with evolutionary algorithms implemented in MATLAB are summarized. An approach was made to use the KEEL Tool, developed in Java by a group of Spanish research centres, to investigate the models. Five fuzzy models comprising 3 or 4 input variables referring to the attributes of a property were learned and evaluated using six regression algorithms for fuzzy rule based systems implemented in the KEEL Tool. The experiments were conducted using a data set prepared on the basis of actual 134 sales transactions made in one of Polish cities and located in a residential section. The results were encouraging. The KEEL Tool has proved to be very useful and effective research tool, especially thanks to its 10-fold cross validation mechanism and relatively short time of data processing.

Key words. real estate appraisal, genetic fuzzy system, cadastral system, KEEL

Introduction

The most popular approach to determining the market value of a property is sales comparison approach. Applying this approach it is necessary to have transaction prices of the properties sold which attributes are similar to the one being appraised. If good comparable transactions are available, then it is possible to obtain reliable estimates. Prior to the evaluation the appraiser must conduct a thorough study of the appraised property using available sources of information such as cadastral systems, transaction registers, performing market analyses, accomplishing on-site inspection. His estimations are usually subjective and are based on his experience and intuition. Automated valuation models (AVMs) are based on statistical models such as multiple regression analysis [3], [18], 24], soft computing and geographic information systems (GIS) [8], [27]. Many intelligent methods have been developed to support appraisers' works: neural networks [10], [26], fuzzy systems [4], [11], case-based reasoning [5], [22], data mining [20] and hybrid approaches [9].

The concept of an evolutionary fuzzy system to assist with real estate appraisals, presented in the paper, was developed basing on sales comparison method [13]. It was assumed that whole appraisal area, that means the area of a city or a district, is divided into sections of comparable property attributes. For each section a representative

property and rule bases should be determined. The architecture of the proposed system is shown in Figure 1. The appraiser accesses the system through the internet and chooses an appropriate section and input the values of the attributes of the property being evaluated. Then the system using the parameters of the representative property for the section indicated, calculates the input values to the fuzzy model. The classic fuzzy inference mechanisms, applying a rule base generated for that section, calculates the output. Then on the basis of the parameters of the representative property the final result is determined and as a suggested value of the property is sent to the appraiser.

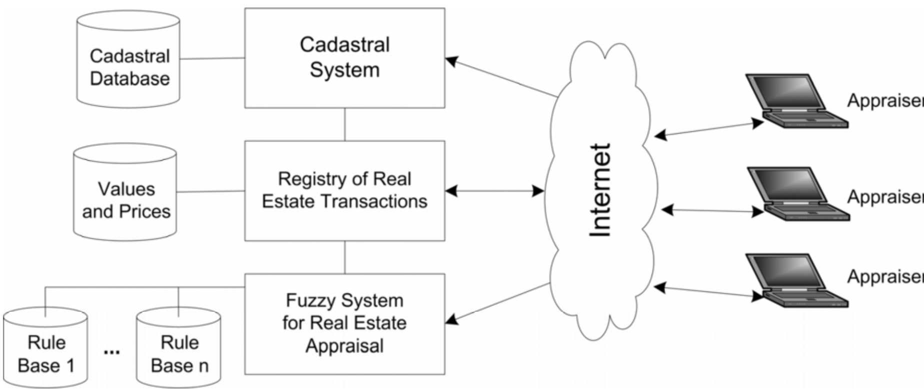


Figure. 1. Information systems to assist with real estate appraisals

Mamdani and Takagi-Sugeno-Kang-type (TSK) fuzzy models were developed and evaluated [14], [15]. The models were built with the aid of experts and comprised 7 input variables relating to main attributes of a property being appraised, namely area, front, infrastructure, neighbourhood, arrangement, distance, and communication. Having constructed the fuzzy model with 7 input variables, the experts were not able to determine a rule base, therefore, an evolutionary algorithm was developed to generate it. The model was implemented using the MATLAB Fuzzy Logic Toolbox and the evolutionary algorithm was programmed to generate the rule base and to tune membership functions, employing the functions included in the MATLAB Genetic Algorithm and Direct Search Toolbox. The set of data used in the experiments concerned sales transactions made in one of Polish cities and located in a residential section what assured comparable attributes of properties.

The experiments were conducted using MATLAB software package were rather time consuming so that we tried to find more time effective solution testing the fuzzy models with reduced number of input variables, different parameters of the rule base, fuzzy membership functions as well as the evolutionary optimization process [14], [15], [16], [17], [21]. In some cases the number of transactions in data sets seemed to be too low to obtain satisfactory results. In some cases the models suffered from overfitting.

We present an approach to use another tool for learning, optimizing and evaluating genetic fuzzy models which is called the KEEL (Knowledge Extraction based on Evolutionary Learning) [1], [2]. The KEEL contains several dozen of algorithms for data pre-processing, designing and conducting the experiments, data post-processing, evaluating and visualizing the results obtained, which have been bound into one flexible and user friendly system. The KEEL has been developed in Java by a group of

Spanish research centres and is available for free for non-commercial purposes (www.keel.es).

1. Outline of Fuzzy Models for Real Estate Appraisal

The fuzzy model for real estate appraisal was built with the aid of experts. The complete model comprised 7 input variables and they referred to the difference or proportion of attribute values between a property being appraised and the representative one. The representative properties were determined for one section of the city comprising residential properties, having similar characteristics, by means of calculating average values of attributes of all properties in the set of data used in the experiment. During our experiments several fuzzy models to assist with real estate appraisal were optimized. The models comprised 2, 3, 4, 5, 6, and 7 input variables which referred to the difference or proportion of attribute values between a property being appraised and the representative one. For each input variable three or five triangular and trapezoidal membership functions were defined the first model variant was denoted as 3FS and the second as 5FS. The linguistic values were as follows: much lower than (MLT), lower than (LT), equal (EQ), greater than (GT), and much greater than (MGT). Thus the inputs of the fuzzy models were defined by vectors composed of following variables:

1. *Area:* is the ratio of the area of the examined parcel to the area of the representative one. The domain of this variable is the interval form 0 to 10. Values greater than 1 indicate that the examined property has the bigger area. The membership functions of the model variants with 3 linguistic values are shown in Figure 2 and with 5 in Figure 3.
2. *Front:* it is the difference in the length of fronts of parcels expressed in meters. The domain of this variable is the interval form -50 to 50 meters. Positive values mean that the examined parcel has longer front than the representative one what is considered as a more valuable case. The membership functions of the model variants with 3 linguistic values are presented in Figure 4 and with 5 in Figure 5.
3. *Arrangement* – pertains to the subjective assessment of how better a given property was arranged than the representative one. Values of this attribute are the appraiser's judgments of what is the difference in this attribute between the appraised and the representative parcel. The values are taken from the range 0-200 where 100 means that both parcels are equal in this respect, values greater than 100 – that the examined parcel is better and the ones lower than 100 – the opposite. The membership functions of the model variants with 3 linguistic values are depicted in Figure 6 and with 5 in Figure 7.
4. *Distance*: it is the difference in the distance from a local centre expressed in meters. The domain of this variable is the interval form -1000 to 1000 meters. Negative values denote that the representative property is located closer to the local centre than the appraised one. The membership functions of the model variants with 3 linguistic values are given in Figure 8 and with 5 in Figure 9.

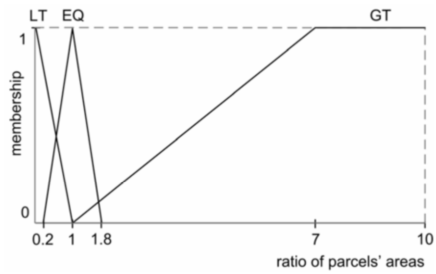


Figure 2. Membership functions of area input variable in 3FS model variant

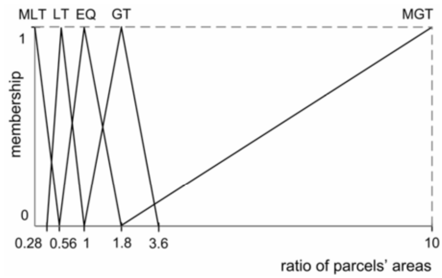


Figure 3. Membership functions of area input variable with 5 linguistic values

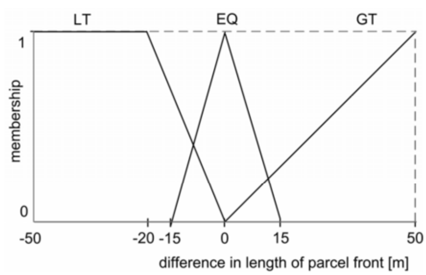


Figure 4. Membership functions of front input variable in 3FS model variant

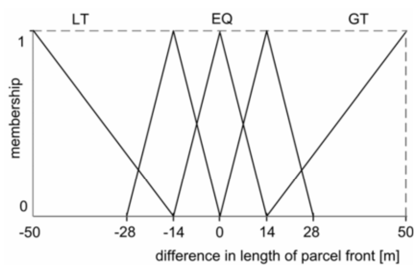


Figure 5. Membership functions of front input variable in 5FS model variant

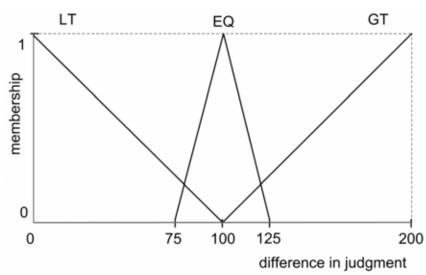


Figure 6. Membership functions of arrangement input variable in 3FS model variant

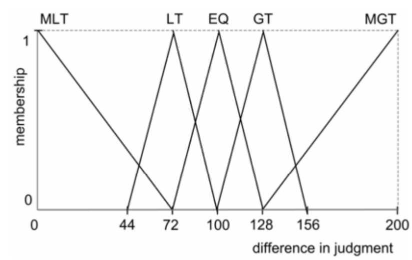


Figure 7. Membership functions of arrangement input variable in 5FS model variant

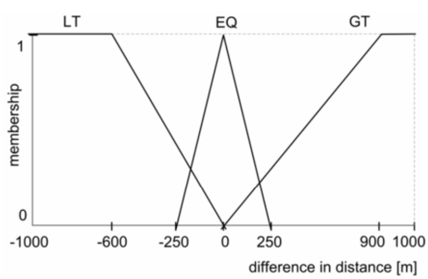


Figure 8. Membership functions of distance input variable in 3FS model variant

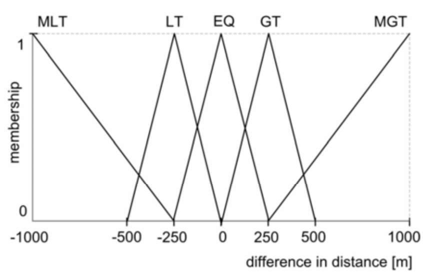


Figure 9. Membership functions of distance input variable in 5FS model variant

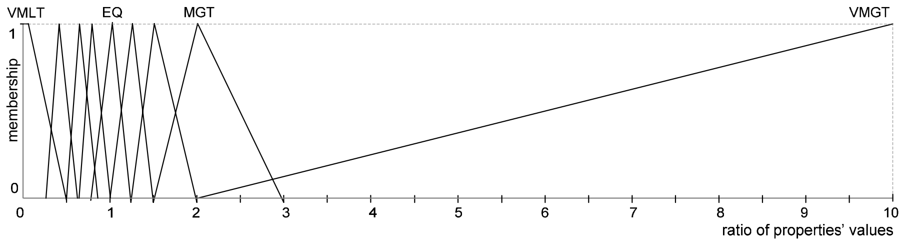


Figure 10. Membership functions of the output of Mamdani-type model

5. *Infrastructure* – refers to what extent the technical infrastructure of a given property is better than the representative one. Values of this attribute are the appraiser's judgments of what is the difference in this attribute between the appraised and the representative parcel. The values are taken from the range 0-200 where 100 means that both parcels are equal in this respect, values greater than 100 – indicate that the examined parcel is better and the ones lower than 100 – the opposite case.
6. *Communication* – deals with the evaluation of the means of public transportation available to the residents. Values of this attribute are the appraiser's judgments of what is the difference in this attribute between the appraised and the representative parcel. The domain of discourse and the meaning of values are the same as in the case of Infrastructure input variable.
7. *Neighbourhood* – concerns the evaluation of the quality of properties' neighbourhood Values of this attribute are the appraiser's judgments of what is the difference in this attribute between the appraised and the representative parcel. The domain of discourse and the meaning of values are the same as in the case of Infrastructure input variable.

Two types of fuzzy models were studied a Mamdani and Takagi-Sugeno-Kang (TSK) ones. The output of the Mamdani-type model was assumed as the ratio of the property value being appraised to the value of the representative one with the domain of discourse from 0 to 10. It was characterized by nine triangular and trapezoidal membership functions (see Figure 10). The linguistic values were as follows: very much lower than (VMLT), much lower than (MLT), lower than (LT), slightly lower than (SLT), equal (EQ), slightly greater than (SGT), greater than (GT), much greater than (MGT) and very much greater than (VMGT). In order to assure that each rule will have influence on the final assessment following operators were used: PROD for aggregation of rule conditions, PROD for activation of rule conclusions and ASUM for accumulation of output membership functions, where PROD means algebraic product and ASUM denotes algebraic sum. The TSK-type model in turn was a zero-order one where output functions were constants representing the difference of values between appraised and representative properties.

Using the complete model with 7 inputs as the base, further models with lower and lower number of input variables were created. As the criterion of eliminating the dimensions the variability coefficient, which is expressed by the standard deviation divided by the mean, was employed. The coefficient was calculated for system input variables using whole set of transaction data (see Table 1). Each successive reduced model comprised input variables with the biggest value of variability coefficient. The

models were denoted by means of codes reflecting the input criteria as shown in Table 2. Some series of experiments were conducted to compare all combinations of three inputs of four with the highest value of variability coefficient as shown in Table 3 [16]. The investigation of the latter five models using KEEL tool is reported in the present paper.

Table 1. Input variables ranked by variability coefficient

No. of variable	Input variable	Standard deviation	Mean	Variability coefficient
1	area	1479.6650	1031.2933	1.4348
2	front	16.8018	21.8333	0.7695
3	arrangement	26.6112	52.3704	0.5081
4	distance	294.4379	580.5333	0.5072
5	infrastructure	17.2505	78.0667	0.2210
6	communication	13.4888	66.3373	0.2033
7	neighbourhood	10.1453	94.6963	0.1071

Table 2. Fuzzy models with different number of input variables evaluated using MATLAB

Model	Input criteria
1234567	area, front, arrangement, distance, infrastructure, communication, neighbourhood
123456	area, front, arrangement, distance, infrastructure, communication
12345	area, front, arrangement, distance, infrastructure,
1234	area, front, arrangement, distance,
123	area, front, arrangement
12	area, front

Table 3. Fuzzy models evaluated using MATLAB and KEEL

Model	Input criteria
1234	area, front, arrangement, distance
123	area, front, arrangement
124	area, front, distance
134	area, arrangement, distance
234	front, arrangement, distance

2. Summary of Previous Experiments Using MATLAB

The models were implemented employing the MATLAB Fuzzy Logic Toolbox and the evolutionary algorithms were programmed to optimize them, using the functions included in the MATLAB Genetic Algorithm and Direct Search Toolbox. Three methods of optimizing the models using evolutionary algorithms were compared. The first one consisted in learning the rule base, the second one in tuning the membership functions having the rule base optimized and the third one in combining both methods in one process. The optimization process of learning the rule base is depicted in Figure 11a and was denoted as the SR process. In this case it was assumed that the membership functions were determined earlier by the expert and were unchangeable. The second method consisted in tuning the parameters of membership functions using the rule base obtained as the product of the previous SR process described above. The optimization process of genetic tuning the membership functions of the fuzzy model is depicted in Figure 11b and was denoted as the SF process. Finally the third optimization process combined both pervious approaches and was denoted as the SM

process. Both the rule base and membership functions were genetically tuned during one process as it was shown in Figure 11c.

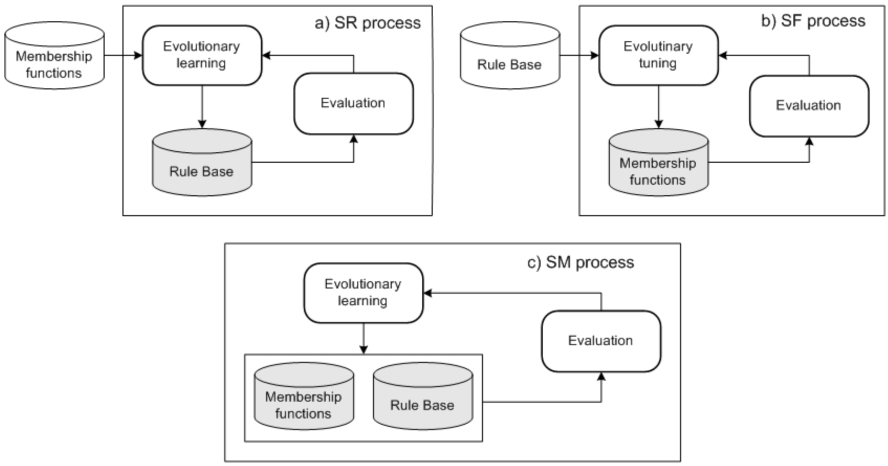


Figure 11. Evolutionary optimizing a) SR process, b) SF process, c) SM process

Training and testing sets. The set of data used in the process of generating rules comprised 134 sales transactions made in one of Polish cities and located in residential sections what assured comparable attributes of properties. The data were taken from the governmental registry of real estate sales transactions. The attributes of the properties embraced by those transactions were determined by an expert, who had visited and studied personally all of them. The set of data was bisected into training and testing sets by clustering the property descriptions including their prices using the k-means method and then by splitting randomly each cluster into two parts. In early experiments the training data set counted 64 properties and the testing one 70 properties. In some late ones the split was made in the proportion of 2:1 and the training and testing data sets comprised 101 and 33 properties respectively.

Coding chromosomes. The rule base was integer and real-coded using the Pittsburgh method, where one chromosome comprised whole rule base. In the SR process the constant length of the chromosome composed of N rules was assumed. Each i-th rule was represented by M+1 genes: g_i^1, g_i^2, …, g_i^{M+1}, where M was the number of input variables and first M genes contained natural numbers from 1 to 3 or 5 corresponding to linguistic values of seven input variables, e.g. MLT (much less than), LT (less than), EQ (equal), GT (greater than), MGT (much greater than) respectively. Zero value on the position of a given input meant that this attribute did not occur in the rule. The M+1th gene represented the output and in the case of the Mamdani type models contained values obtained in similar way as the values of input genes. In the TSK-type models in turn, the output gene contained natural numbers which were drawn at random from the set of successive numbers from 1 to 81 representing the difference of values between appraised and representative properties expressed in Polish currency (PLN). The difference could range from -200 to 200 PLN with the step of 5 PLN, what established altogether 81 values. In the SF process chromosomes contained only the representation of membership functions. In turn in the SM process the chromosome

was constructed by attaching the representation of membership functions to the one reflecting the rule base.

Fitness function was calculated as the mean relative error (MRE) between prices of properties included in the training set and the prices of corresponding properties predicted by the fuzzy system using a rule base produced by a subsequent generation of the evolutionary algorithm.

Genetic operations. The parameters of reproduction were as follows: elite count was set to 2, crossover fraction was set to 0.8, and therefore the mutation fraction was close to 0.2. Uniform crossover operation was employed, where the pattern of the position of rules to be exchanged was determined randomly for each pair of parents separately with the probability of 0.5. According to this pattern whole rules were exchanged between the parents of a given pair instead of individual genes. In the SM process chromosome fragments representing rule base and membership functions were processed separately. The mutation operation consisted in altering rules randomly selected with a given probability in each chromosome, remained in the mutation fraction. In the SM process chromosome fragments representing rule base and membership functions were mutated separately.

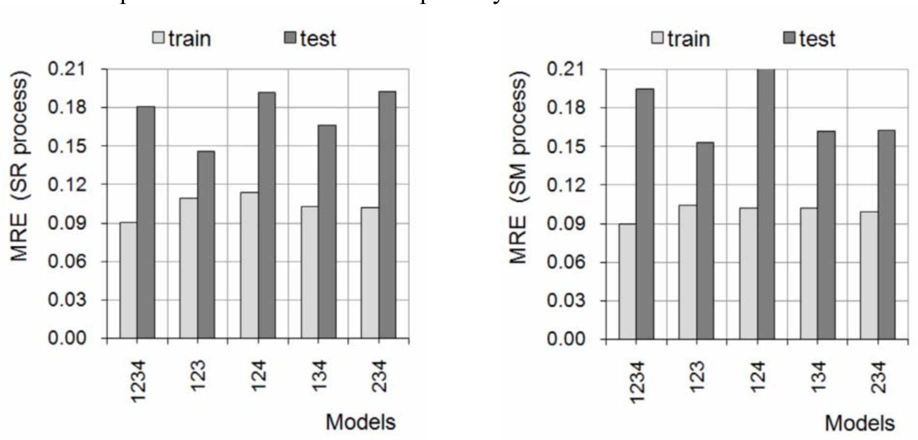


Figure 12. Results of fuzzy model optimization with SR and SM processes in MATLAB

3. Fuzzy Rule Learning Algorithms Implemented in KEEL Tool

Keel is a non-commercial Java software tool [1], [2] designed to assess evolutionary algorithms for Data Mining problems. It enables the researches to solve regression, classification, clustering, pattern mining problems. Genetic fuzzy algorithms based on different approaches such as Pittsburgh, Michigan, IRL, GCCL are encapsulated into one system. KEEL is done for different users with different expectations and provides three main functionalities: *Data Management*, which is used to set up new data, import, export data in other formats to KEEL format, data edition and visualization, applying new transformations and partitioning data etc., *Experiments*, which is used to design and evaluate experiments with use of selected data and provided parameters, *Educational*, which is used to design experiment and run it step-by-step in order to display learning process.

KEEL has the following main features. It presents algorithms in predicting models (pre-processing and post-processing). It offers pre-processing algorithms such as data transformation, discretization, instance selection, feature selection, missing values supplement. Post-processing algorithms can be used to refine the results obtained by knowledge extraction algorithms and they ensure membership function tuning, fuzzy rule weighing and selection. It also contains a statistical library to analyze and compare the results of the algorithms. Moreover it provides on-line experimentation for educational purposes and off-line experimentation for research purposes. User-friendly graphical interface enables the researchers to design experiments by joining individual functions and components into whole processes. The graph of the experiment reported in the present paper is shown in Figure 13.

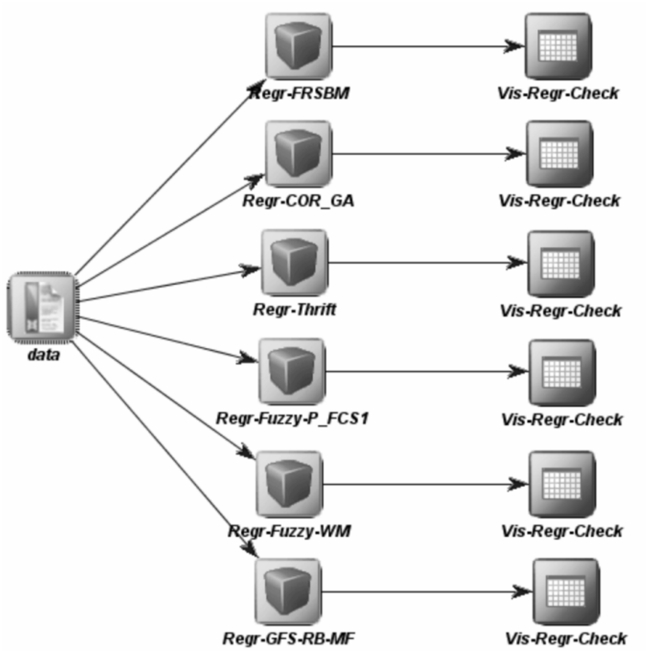


Figure 13. Graph of the experiment created in KEEL

The great advantage of KEEL to is the possibility to create different experimental sets of data automatically and to perform cross validation of learned models within the same process, what substantially decreases time needed to prepare and accomplish investigations and makes possible to avoid or diminish the threat of model overfitting.

Six following KEEL algorithms for genetic learning of fuzzy rules were employed to carry out experiments with the same data which were used in pervious tests using MATLAB:

COR_GA. Proposed by Casillas Cordon Herrera [7] the cooperative rules methodology is an approach to generation simple, accurate linguistic fuzzy rules. It is an improvement for the ad hoc data-driven methods. Instead of choosing the consequents with highest performance in current subspaces this methodology investigates other consequents, which differ from the best one, to be chosen. The aim is to make FRBS more accurate in order to find the best cooperation in KB. First search space is defined

by obtaining candidate rules for each fuzzy input subspace. Then the most cooperative fuzzy rule set is performed by the combinatorial search. Algorithm is evaluated among candidate rules searching for these consequents which reach the best global accuracy.

FRSBM. It is the random sets-based method proposed by Sanchez [19], to build up a fuzzy model of imprecise measurements taken on a physical system. Missing inputs of the model are reflected by means of a random - measurements are imprecise, inexact. Then variables observed in such environment are treated as experiments, which are modelled (as, with, by) fuzzy random variables. The level cuts (random sets) of the variables determine their interpretation. Fuzzy model is built from memberships of fuzzy sets that are compliant to one-point coverage of random sets. Procedure to adjust parametric membership subsets of the initial samples was proposed. Uncertainty-based quality measure of the fuzzy models is foundation to select subsets. The aim is to reach the precision and simple model.

GFS-RB-MF. An algorithm developed by Homaifar, McCormick [12]. It is based on Simple Genetic Algorithm, which is used to optimized parameters of the fuzzy controller. It aims to design of membership functions and rule sets for fuzzy controller in parallel. Population size, maximum number of generations, probability of crossover and mutation are provided by a user. While traditional Simple Genetic Algorithm is binary-based, this method is integer-based and optimizes whole knowledge base of the system. The size of the rule set and added to the number of fuzzy sets of the input and output. Individual locations which build the string so called alleles describe rule set and the parameters of the membership function. Every possible combination of input fuzzy sets is included in the rule set. Membership functions are encoded by the values from a given range. Thus, in one string rule base and membership function are encapsulated.

P_FCS1. The method was presented by Carse, Fogarty, Munro [6]. It is based on Pittsburgh model and is an approach to genetics-based reinforcement learning of fuzzy controllers. Genetic operations are performed at the level of the complete fuzzy rule-set, number of rules in each chromosome may differ. Genetic algorithm, which includes rule base and fuzzy membership functions (encoded with real numbers) optimization, is the tool/mechanism used. Rules are encoded in the same manner as in the Parodi's and Bonelli's approach. For each variable in the condition and action part there is assigned their fuzzy set membership function:
$R_i(x_{c1i}, x_{w1i});\dots(x_{cni}, x_{wni}) \to (y_{c1i}, y_{w1i});\dots(y_{cni}, y_{wni})$, where x_{cni}, y_{c1i} defines centre of membership function and x_{wni}, y_{w1i} defines width of membership function.
Due to specific encoding genes, modified genetic operators are provided. Chosen random value is used to determine the cross point according to the centres of the input membership functions assigned to each rule.

WM. A method worked out by Wang and Mendel [25]. It strives to provide a common fuzzy rule base from fuzzy rules which are generated from the data pairs (examples) and the linguistic fuzzy rules. It is divided into four steps. First variable spaces are divided into fuzzy regions (each domain interval into 2N+1 regions). Then the fuzzy rules are generated from given data pairs. Each rule variable is set to the region with the maximum degree. The third step considers assigning a degree to each rule. Firing levels of each variable in rule are multiplied, as described by following example:
$D(Rj)=\mu A1(x_1)^{*}\dots\mu A1(x_1)^{*}\mu B1(y)$, where rule Rj is described as follows:
IF(x_1 is A_1) I (x_2 is A_2) ... I (x_n is A_n) THEN (y is B)

The last step encloses determining of the rule base. Rules are partitioned according to their antecedents. The rule with the highest degree is chosen from each group.

Thrift. A method proposed by Thrift [23] method. It is based on the Pittsburgh approach and comprises learning fuzzy rules only, with a fixed set of fuzzy MFs set by hand. Rule base is represented by a relational matrix. The rules are encoded into the chromosome while fixing the membership function.

4. Results of the Experiment

An attempt was made to carry out experiments in KEEL using our set of data comprising 134 property attributes and sales transactions. All algorithms, i.e. COR_GA, FRSBM, GFS-RB-MF, P_FCS1, WM and Thrift's one, accepted n inputs and one output so that they fully conformed to data we possessed. Five fuzzy models comprising 1234, 123, 124, 134, 234 input variables were learned using these algorithms. Parameters of the algorithms were remained with their default values. The results obtained for successive models for 10-fold cross validation are shown in Figure 14-18. Two measures of output quality were used. First one was the default measure implemented in KEEL i.e. mean square error (MSE). As the second measure we applied the root mean square error divided by average price (RMSE/avg) in order to obtain non-dimensional values which could be compared with mean relative error (MRE) used in previous experiments with MATLAB.

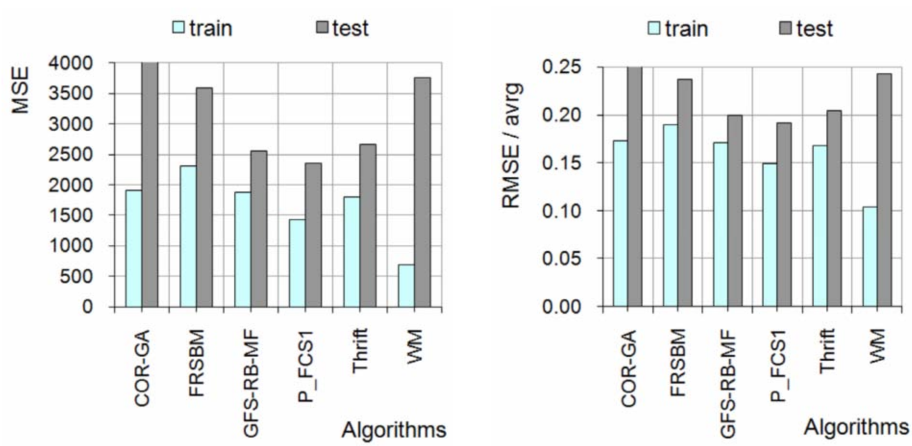


Figure 14. Results for 1234 input variables and 6 regression algorithms

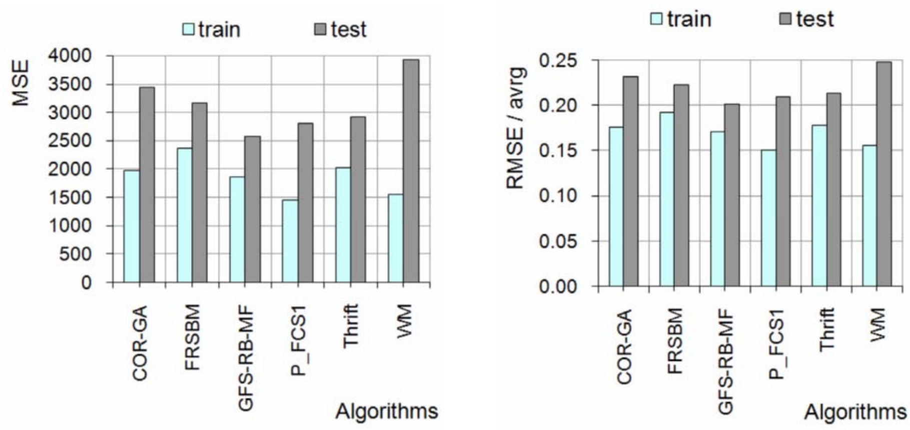


Figure 15. Results for 123 input variables and 6 regression algorithms

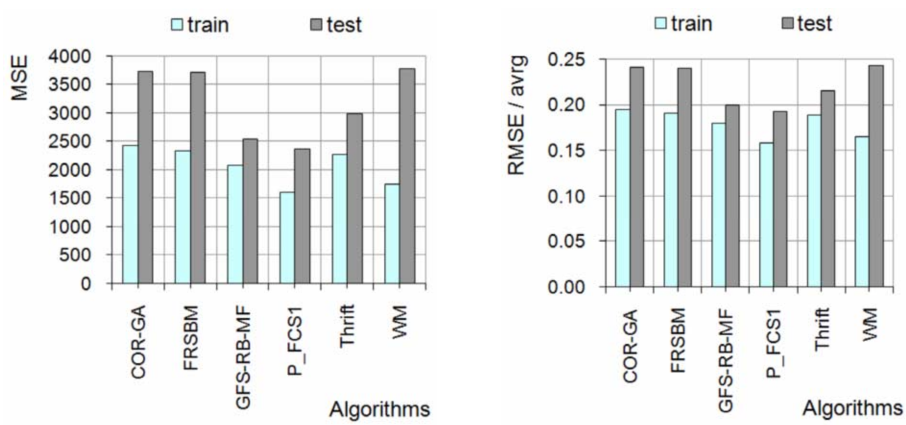


Figure 16. Results for 124 input variables and 6 regression algorithms

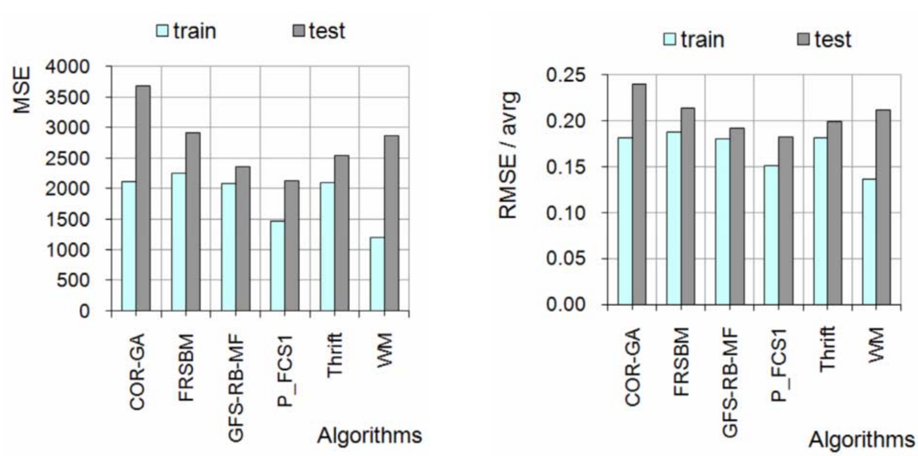


Figure 17. Results for 134 input variables and 6 regression algorithms

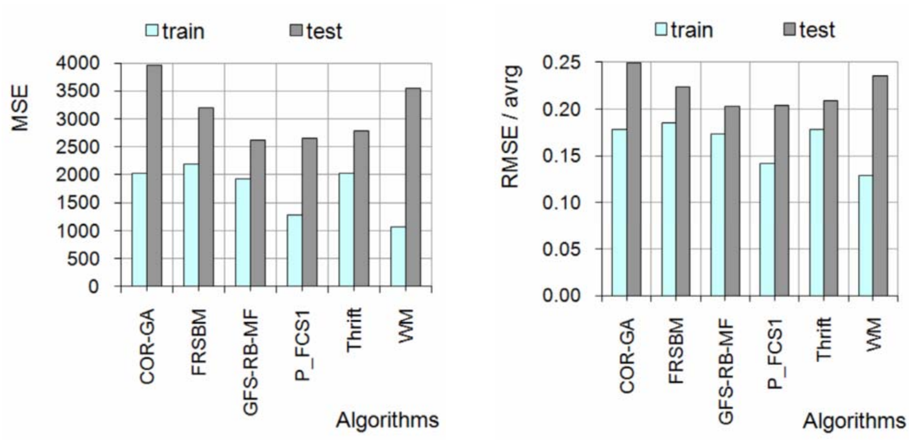


Figure 18. Results for 234 input variables and 6 regression algorithms

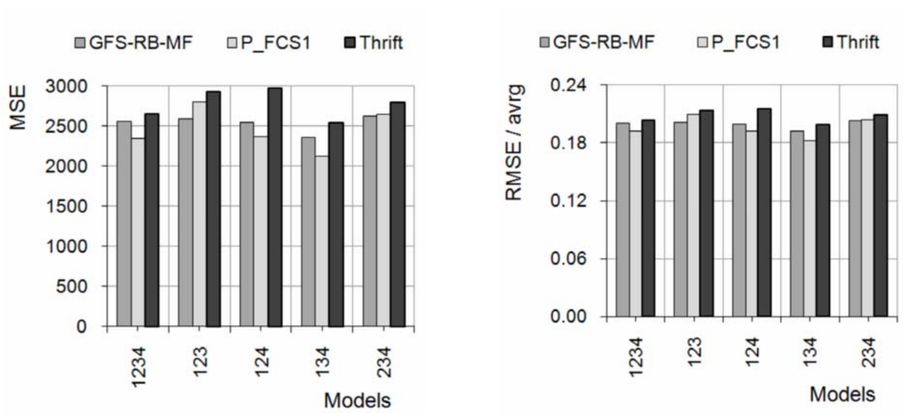


Figure 19. Results of model comparison for selected algorithms and testing data

It can be easily seen that GFS-RB-MF, P_FCS1 and Thrift's algorithms provided the best results. However the conclusion cannot be decisive because there was no parameter tuning. Comparative results for these three algorithms for all five models were presented in Figure 19. One can notice that the 134 model has the best performance.

5. Conclusions and Future Work

Five fuzzy models comprising 1234, 123, 124, 134, 234 input variables were learned using the COR_GA, FRSBM, GFS-RB-MF, P_FCS1, WM and Thrift's algorithms implemented in the KEEL Tool. The experiments were conducted using data prepared on the basis of actual 134 sales transactions of residential properties. 10-fold cross validation of the models learned was set. The results of these initial experiments are encouraging. The values of RMSE/avrg measure for testing data were about 0.2, what

is somewhat more when comparing with the optimization using Matlab, but no tuning parameters of KEEL algorithms was made.

The KEEL Tool has proved to be very useful and effective research tool, especially thanks to 10-fold cross validation mechanism and relatively short time of data processing in the case of our tests. Further experiments with different parameters of learning algorithms and with pre- and post-processing are planned. Moreover out research team has obtained a much bigger set of sales transactions data concerning apartments, which will constitute the basis of further investigations.

References

[1] J. Alcala-Fdez, L. Sanchez, S. Garcıa, M.J. del Jesus, S. Ventura, J.M. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas, J.C. Fernandez, F. Herrera, *KEEL: A Software Tool to Assess Evolutionary Algorithms for Data Mining Problems*, 2008, http://www.keel.es/ (to be published in Soft Computing).

[2] J. Alcala-Fdez, S. Garcıa, F.J. Berlanga, A. Fernandez, L. Sanchez, M.J. del Jesus, F. Herrera, KEEL: A data mining software tool integrating genetic fuzzy systems. *Proceedings of the 3rd International Workshop on Genetic and Evolving Fuzzy Systems (GEFS'08*), F. Hoffmann et al., eds, IEEE, Piscataway, NJ, 2008, 83-88.

[3] M. d'Amato, Comparing Rough Set Theory with Multiple Regression Analysis as Automated Valuation Methodologies, *International Real Estate Review* 10(2) (2007), 42–65.

[4] C. Bagnoli, H. C. Smith, The Theory of Fuzzy Logic and its Application to Real Estate Valuation, *Journal of Real Estate Research* 16(2) (1998), 169–199.

[5] P. P. Bonissone, W. Cheetham, Financial Applications of Fuzzy Case-Based Reasoning to Residential Property Valuation, *In Proc. of the 6th IEEE International Conference on Fuzzy Systems, Barcelona*, 1997, 37-44.

[6] B. Carse, T.C. Fogarty, A. Munro, Evolving fuzzy rule based controllers using genetic algorithms, *Fuzzy Sets and Systems* 80:3 (1996) 273-293.

[7] J. Casillas, O. Cordón, F. Herrera, COR: A methodology to improve ad hoc data-driven linguistic rule learning methods by inducing cooperation among rules, *IEEE Transactions on System, Man and Cybernetics, Part B: Cybernetics* 32:4 (2002) 526-537.

[8] G.H Castle, GIS: Meeting the Information Demand, *Valuation Insights and Real Estate Investor* 42(1) (1998) 66–71.

[9] W.J. McCluskey, S. Anand: The application of intelligent hybrid techniques for the mass appraisal of residential properties, *Journal of Property Investment and Finance*, Vol.17, No.3, (1999) 218-239.

[10] Q. Do, G. Grudnitski: A Neural Network Approach to Residential Property Appraisal. *Real Estate Appraiser*, December 1992, 38–45.

[11] M.A.S. González C.T. Formoso: Mass appraisal with genetic fuzzy rule-based systems, *Property Management* 24(1) (2006), 20-30.

[12] A. Homaifar, E. McCormick: Simultaneous Design of Membership Functions and Rule Sets for Fuzzy Controllers Using Genetic Algorithms, *IEEE Transactions on Fuzzy Systems* 3:2 (1995) 129-139.

[13] D. Król, T. Lasota, W. Nalepa, and B. Trawiński: *Fuzzy system model to assist with real estate appraisals, Lecture Notes in Artificial Intelligence* 4570, 2007, 260–269.

[14] D. Król, T. Lasota, B. Trawiński, and K. Trawiński: Comparison of Mamdani and TSK Fuzzy Models for Real Estate Appraisal, *Lecture Notes in Artificial Intelligence* 4693, 2007, 1008–1015.

[15] D. Król, T. Lasota, B. Trawiński, and K. Trawiński: Investigation of evolutionary optimization methods of TSK Fuzzy Model for Real Estate Appraisal, *International Journal of Hybrid Intelligent Systems* 5 (2008) 1–18.

[16] T. Lasota, B. Trawiński, and K. Trawiński: Evolutionary Generation of Rule Base in TSK Fuzzy Model for Real Estate Appraisal, *Proceedings of the 3rd International Workshop on Genetic and Evolving Fuzzy Systems* (GEFS'08), F. Hoffmann et al., eds, IEEE, Piscataway, NJ, 2008, 71-76.

[17] T. Lasota, B. Trawiński, and K. Trawiński: Evolutionary Optimization of TSK Fuzzy Model to Assist with Real Estate Appraisals, in *Computational Intelligence: Methods and Applications*, L. Rutkowski, R. Tadeusiewicz, L.A. Zadeh and J. Zurada, eds, EXIT, Warszawa, 2008, 232-243.

[18] N. Nguyen, A. Cripps: Predicting housing value: A comparison of multiple regression analysis and artificial neural networks, *Journal of Real Estate Research* 22(3) (2001), 3131–3336.

[19] L. Sánchez,: A Random Sets-Based Method for Identifying Fuzzy Models, *Fuzzy Sets and Systems* 98:3 (1998) 343-354.

[20] L. Soibelman, M.A.S. González: A Knowledge Discovery in Databases Framework for Property Valuation, *Journal of Property Tax Assessment and Administration*, Vol. 7, No. 2, (2002) 77-106.

[21] M. Sygnowski, B. Trawiński, and A. Zgrzywa,: An Attempt to Use a Type-2 Fuzzy Logic System to Assist with Real Estate Appraisals, *Proceedings of the 1st International Conference on Information Technology* (IT 2008), A. Stepnowski et al., eds, IEEE, Piscataway, NJ, 2008, pp. 189-192.

[22] W.Z. Taffese,: Case-based reasoning and neural networks for real state valuation, *Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi-Conference: artificial intelligence and applications*, Innsbruck, Austria (2007).

[23] P. Thrift: Fuzzy logic synthesis with genetic algorithms, *Proceedings of 4th International Conference on Genetic Algorithms* (ICGA'91), pp 509–513.

[24] B.D. Waller, T.H. Greer, N.F. Riley: An Appraisal Tool for the 21st Century: Automated Valuation Models, *Australian Property Journal* Vol. 36 No. 7 (2001) 636–641.

[25] L.X. Wang, J.M. Mendel: Generating Fuzzy Rules by Learning from Examples, *IEEE Transactions on Systems, Man and Cybernetics* 22:6 (1992) 1414-1427.

[26] E. Worzala, M. Lenk, A. Silva: An Exploration of Neural Networks and Its Application to Real Estate Valuation, *The Journal of Real Estate Research* 10(2) (1995), 185–201.

[27] P. Wyatt: The development of a GIS-based property information system for real estate valuation, *International Journal of Geographical Information Science*, Vol. 111, No. 5 (1997) 435–450.

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-140

Optimization of Top-k Spatial Preference Queries' Execution Process Based on Similarity of Preferences

Marcin GORAWSKI and Kamil DOWLASZEWICZ
Institute of Computer Science, Silesian University of Technology, Poland

Abstract. A detailed description of top-k spatial preference queries and the schema of their execution will be presented. Then, it discusses existing R-tree based top-k spatial preference queries' execution algorithms and the optimization methods they utilize. Moreover, the paper presents a new optimization technique and an algorithm capable of utilizing it together with other methods. Finally, an analysis of proposed method's efficiency is presented.

Keywords. Top-k spatial preference queries, optimization, similar preferences

Introduction

Top-k spatial preference queries retrieve objects having greatest value in terms of chosen attributes describing other objects from their neighbourhood [1]. Therefore, the data they operate on is described by location and other, non-spatial attributes. The queries can be utilized in various decision support systems. They can be applied in such domains like service recommendation, investment planning, or emergency management.

1. Queries

Top-k spatial preference query specifies target data set and attributes of feature objects which can influence the value of target objects. It also specifies a method of choosing objects which will be used to compute the target objects' value. These can for example be their nearest neighbours or objects located within some specified distance from them. The latter requires further parameterization of the query; the method of computing target object's value depending on attribute values of objects found in its neighbourhood has to be chosen. It can for example return maximum, minimum or an average of its neighbouring feature objects' attributes. As each target objects' value depends on the query, it is therefore subjective [1].

An example of top-k spatial preference query is a search for residential buildings located near forests of high fire risk and warehouses storing dangerous materials.

Figure 1 illustrates the execution of such query. White circles represent target objects which in this case are residential buildings. Black circles represent warehouses and gray rectangles represent forests; depending on their chosen attribute values target

objects' value will be computed. However, each target object is influenced only by feature objects belonging to its neighbourhood.

The query execution schema presented on the figure concerns two cases where preferences based on both forest and warehouse set use the same method defining the neighbourhood. When neighbourhood is defined as a region within some specified range from the target object, it is presented on the figure as dashed circle. This method of neighbourhood definition is further called range method. On the other hand, when neighbourhood is defined as the object least distant from the target object it is represented by a solid line. This neighbourhood definition method is further called nearest neighbour method.

In the case of range method, assuming that the function computing target objects' partial values returns maximum of neighbouring feature objects' attribute values, object r1 is found. Its total value is highest and amounts to 0.8+0.6=1.4, while r2 object's value equals 0.3+0.9=1.2, r3 object's value equals 0.7+0=0.7 and r4 object's value equals 0+0.6=0.6.

In the case of nearest neighbour method, r3 is the object of highest value. Its total value amounts to 0.7+0.9=1.6, while r1 object's value equals 0.8+0.6=1.4, r2 object's value equals 0.3+0.9=1.2 and r4 object's value equals 0.3+0.6=0.9.

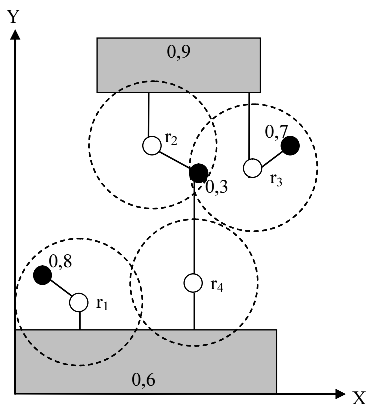


Figure 1. A schema of query execution using range and nearest neighbour neighbourhood definition methods

A different example of top-k spatial preference query is a search for apartments having in its vicinity vegetarian restaurants offering a wide selection of food and convenient access to public transport.

2. Query Execution Process

Every spatial preference query consists of numerous typical spatial queries like range and nearest neighbour search. This query class is also characterized by operating on data sets containing numerous objects. A search for objects in terms of their spatial location can be effectively realized using spatial indices. This text focuses on methods using R-tree [2] which is the most popular spatial index.

An R-tree is a dynamic, balanced tree structure, which allows efficient execution of typical spatial queries like nearest neighbour and range search [2]. The queries can

be realized both on objects whose spatial location is defined as a point and objects covering a specified area. Minimum bounding rectangles estimating the area covered by each object are stored in the structure at the leaf level together with the pointer to the indexed object. The higher level nodes store entries with a minimum bounding rectangle covering the bounding rectangles of their children and with pointers to these child nodes. As a result nodes at the higher level represent larger areas, while their children represent smaller, more precise ones [3].

During target object's value computation two R-tree backed basic spatial queries can be executed.

The range query finds objects having at least one common point with the specified region [2]. Figure 2 presents an example of such query. It searches for objects located within distance of 2 from object q. The search region is represented by dashed circle surrounding object q. Starting from the tree root the query process recursively visits node's children whose minimum bounding rectangles intersect the query region. In the discussed example e1 entry does not intersect the query region and is hence discarded. Entry e2 intersects the query region therefore the algorithm is recursively executed on it. During that process minimum bounding rectangles of o1 and o7 objects' are found to intersect the query region. Assuming that the objects in fact intersect the query region, they are added to the result set. Analogously to e2, entry e3 does not intersect the query region and its branch can be discarded.

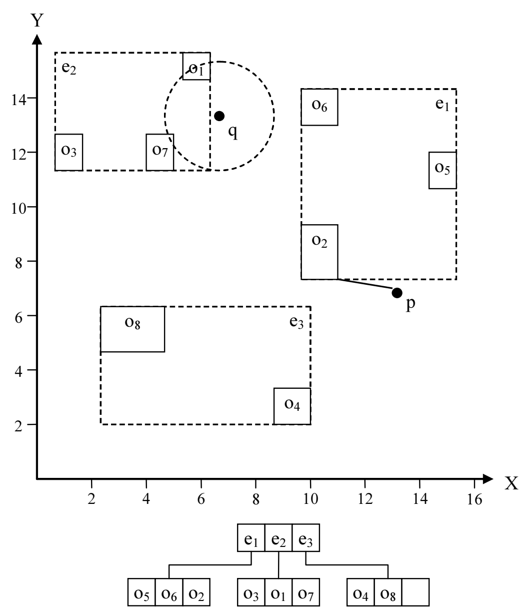


Figure 2. R-tree backed search for objects located in the neighbourhood of target object

Nearest neighbour query, on the other hand, finds objects least distant from the query region. An example of such query is presented in figure 2 and searches for objects being nearest neighbours of object p. In order to realize such a query a best-first [4] algorithm can be utilized. It stores the entries in a priority queue ordered by their distance from object p. The algorithm starts at the tree root and inserts into the queue entries e1, e2, and e3. Then the least distant entry is removed from the queue and

analyzed. This results in removing e1 entry and inserting its child o5, o6, and o2 entries. Next entry removal from the queue results in finding object o2 which is the nearest neighbour of p.

3. Algorithms

The most straightforward method to realize the query is computing the value of all target objects and choosing best-k amongst them. This approach however has a significant drawback. It requires computing every partial value of every target object, which in turn requires executing a spatial query for all preferences for every target object. This results in a large number of index and data accesses.

3.1. Discarding Objects

The SP algorithm presented in [1] assumes that the partial value based on every preference can always be represented within [0,1] range. Then, it is possible to define the maximum possible value of the target object, for which not every partial value is yet known. In [1] it is defined as:

$$t_{+}^{\theta}(p) = agg_{c=1}^{m} \begin{cases} t_{c}^{\theta}(p) \text{ when } t_{c}^{\theta}(p) \text{ is known} \\ 1 \text{ otherwise} \end{cases} \tag{1}$$

It allows optimizing the query execution by discarding objects whose maximum possible values are lesser than the total value of the k-th best object at that moment. The computation of their value is being stopped. This reduces the number of executed spatial queries and, in consequence, lowers the number of index and data accesses.

3.2. Simultaneous Value Computation

It is possible to further optimize the process by utilizing R-tree specifics. It stores data at the leaf level and the relative distance between objects in one R-tree leaf, located in one region, is short. Simultaneous computation of these objects' values should reduce the number of accesses, as the feature objects belonging to their neighbourhood will also belong to the same or neighbouring leaves of feature object index. This trait is utilized by the GP algorithm [1].

3.3. Similar Preferences

The optimization based on similarity of preferences is a new top-k spatial preference query optimization method. It can be utilized when query contains any preferences that fulfil specified conditions. They have to be based on one feature data set and define neighbourhood analogously. The first condition guarantees that partial values based on the preferences are computed on the basis of objects found through a spatial search performed on one index. The second condition guarantees that the search area is analogous and therefore the same objects will be found. Preferences which meet these conditions will be further called similar.

When preferences are similar, partial values based on them can be computed during one feature index search. It allows reducing the number of accesses made during query execution. For example any two preferences defining neighbourhood using range method and specifying the same value of range parameter are regarded as similar.

The idea of the proposed technique is graphically presented on figure 3, where node entries intersecting the neighbourhood region are coloured gray.

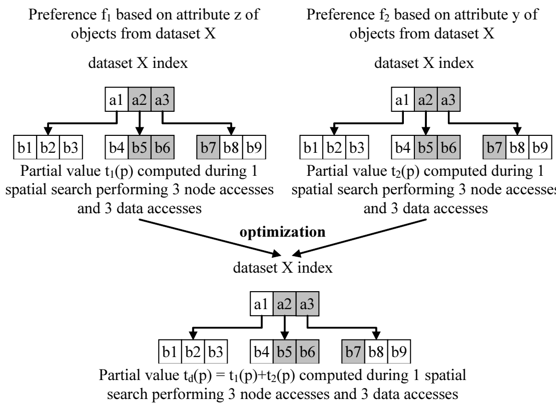


Figure 3. The concept of optimization based on similarity of preferences

The proposed technique also performs an additional operation which further optimizes query execution when used together with object discarding technique. Partial values based on sets of similar preferences are computed earlier than values based on single preferences. Analogical order is ensured for sets of similar preferences. Those of higher cardinality are given higher priority. This results in computing the components being the biggest fraction of total object's value earlier and allows us to discard objects after a lower number of index and data accesses.

Figure 4 presents an example of a query consisting of seven preferences. Four of them marked by A are mutually similar. Preferences marked by B are also similar. Assuming that all partial values of object p equal 0, its value is very low and it should be quickly discarded. However, the number of spatial searches, after which it will happen, depends on the order in which the partial values will be computed.

The benefit of simultaneous computation of partial values based on similar preferences and the benefit of computing partial values in specific order according to the cardinality of similar preferences sets can be observed by analyzing table 1. It presents the minimum value of current k-th best object for object p to be discarded when partial values are computed in different order.

When all partial values are computed individually, the discarding value has to be relatively high in order to discard an object. For example assuming the discarding value equals 3.5 and the proposed optimization technique is not used, at least four spatial searches have to be executed for the object to be discarded. When, however, the proposed technique is used, partial values based on similar preferences are not only computed simultaneously, but also in optimized order, marked by bold font. This results in only one spatial search to be necessary before discarding object p.

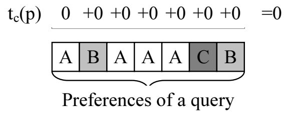


Figure 4. A graphical concept of a query consisting of seven preferences and an example of object *p*, which all partial values equal zero

Table 1. Minimum discarding value for object *p* to be discarded after specific number of spatial queries

	Minimum total value of currently k-th best object for object p to be discarded after consecutive spatial searches.						
Order of computing partial values	1 spatial search	2 spatial searches	3 spatial searches	4 spatial searches	5 spatial searches	6 spatial searches	7 spatial searches
ABC	**>3**	**>1**	**>0**	**N/A**	**N/A**	**N/A**	**N/A**
ACB	>3	>2	>0	N/A	N/A	N/A	N/A
BAC	>5	>1	>0	N/A	N/A	N/A	N/A
BCA	>5	>4	>0	N/A	N/A	N/A	N/A
CAB	>6	>2	>0	N/A	N/A	N/A	N/A
CBA	>6	>4	>0	N/A	N/A	N/A	N/A
Similar preferences technique not used	>6	>5	>4	>3	>2	>1	>0

4. Implementation

During experiments, an application utilizing R*-tree [5] was used. It allowed executing top-k spatial preference queries using various algorithms. A list of all implemented algorithms and their corresponding optimization methods is presented in table 2.

Table 2. Optimization techniques utilized by implemented query algorithms

	Simultaneous computation of partial values of objects from one R-tree leaf technique	Object discarding technique	Similar preferences technique
L0P0S0	no	no	no
L0P1S0	no	yes	no
L0P0S1	no	no	yes
L0P1S1	no	yes	yes
L1P0S0	yes	no	no
L1P1S0	yes	yes	no
L1P0S1	yes	no	yes
L1P1S1	yes	yes	yes

One of the operations realized during top-k spatial preference query execution is the computation of target object's value. It is realized by finding feature objects belonging to the target object's neighbourhood and examining the value of their chosen attributes. For implementation reasons and in order to preserve clarity algorithms computing values of target objects were implemented separately. Their pseudo code is presented as algorithms 1, 2, 3, and 4. Algorithms executing top-k spatial preference queries computing each object's value individually utilize algorithms RNGValue and NNValue. Algorithms GroupRNGValue and GroupNNValue are utilized by top-k spatial preference query algorithms computing many objects' values simultaneously. Algorithms computing target objects' values were designed on the basis of their general description presented in [1] and extended by the mechanism of simultaneous computation of partial values based on similar preferences. Algorithms NNValue and GroupNNValue also drew on best-first nearest neighbour algorithm presented in [4].

Algorithm 1 RNGValue. Computes object's partial value using range method

```
RNGValue(node of index storing feature objects N, target object p, distance d, similar
preferences set S)
1.   For every entry e ∈ N:
2.        If N is a leaf:
3.             If distance between e entry's MBR and p ≤ d:
4.                  Get object g pointed by e.
5.                  If distance between g and p ≤ d:
6.                       For every preference f_c ∈ S: [simultaneous computation of object's
                         partial values based on similar preferences]
7.                            Compute partial value t_c(p) on the basis of object's g feature's c
                              value specified by f_c and current value of t_c(p)
8.        Otherwise:
9.             If distance between e entry's MBR and p ≤ d:
10.                 Execute RNGValue(e, p, d, S).
```

Algorithm 2 NNValue. Computes object's partial value using nearest neighbor method

```
NNValue(tree root of index storing feature objects R, target object p, similar preferences set S)
1.   Create a priority queue Q storing entries ordered by their distance from p.
2.   Insert R into Q.
3.   Repeat until nearest neighbor is found or Q is empty:
4.        Remove the first element e from Q.
5.        If e is a feature object:
6.             For every preference f_c ∈ S: [simultaneous computation of object's partial values
               based on similar preferences]
7.                  Compute partial value t_c(p) on the basis of object's g feature's c value
                    specified by f_c and current value of t_c(p)
8.             If current first entry in Q is more distant from p than e:
9.                  Mark the fact of founding the nearest neighbor.
10.       Otherwise if e is a leaf node:
11.            Insert into Q the object pointed by e.
12.       Otherwise if e is an inner tree node:
13.            For every entry y ∈ e:
14.                 Insert y into Q.
```

Algorithm 3 GroupRNGValue. Simultaneously computes objects' partial values using range method

```
GroupRNGValue(node of index storing feature objects N, a set of target objects V, distance d,
similar preferences set S)
1.   For every entry e ∈ N:
2.        If N is a leaf:
3.             For every object p from V: [simultaneous computation of partial values of objects
               from one leaf]
4.                  If distance between e entry's MBR and p ≤ d:
5.                       Get object g pointed by e.
6.                       If distance between g and p ≤ d:
7.                            For every preference f_c ∈ S: [simultaneous computation of object's
                              partial values based on similar preferences]
8.                                 Compute partial value t_c(p) on the basis of object's g
                                   feature's c value specified by f_c and current value of t_c(p)
9.        Otherwise:
10.            For every p ∈ V:
11.                 If distance between e entry's MBR and p ≤ d:
12.                      Execute GroupRNGValue(e, V, d, S)
13.                      Exit the innermost loop
```

Algorithm 4 GroupNNValue. Simultaneously computes objects' partial values using nearest neighbor method

```
GroupNNValue(tree root of index storing feature objects R, a set of target objects V, similar
    preferences set S)
1.  On the basis of spatial location of objects belonging to V create centroid C
2.  Create T set storing target objects for which nearest neighbor is not found and insert into
    it all objects p ∈ V.
3.  Create a priority queue Q storing entries ordered by their distance from C
4.  Insert R into Q
5.  Repeat until T = Ø or Q = Ø:
6.      Remove the first element e from Q.
7.      If e is a feature object:
8.          For every object p ∈ T: [simultaneous computation of partial values of objects
            from one leaf]
9.              If currently known closest distance between p and a feature object ≥ distance
                between p and e:
10.                 Update the closest distance between p and a feature object using the
                    distance between p and e.
11.                 For every preference f_c ∈ S: [simultaneous computation of object's
                    partial values based on similar preferences]
12.                     Compute partial value t_c(p) on the basis of object's g feature's c
                        value specified by f_c and current value of t_c(p).
13.             If the distance between current first entry in Q and C > sum of distances
                between p and C and current closest distance between p and a feature object:
14.                 Remove p from T.
15.     Otherwise if e is a leaf node:
16.         Insert into Q the object pointed by e.
17.     Otherwise if e is an inner tree node:
18.         For every entry y ∈ e:
19.             Insert y into Q.
```

Top-k spatial preference query algorithm L1P1S1, presented as algorithm 5, utilizes object discarding, simultaneous computation of many objects' values and the presented technique based on similar preferences. This uses algorithms GroupRNGValue and GroupNNValue to compute partial values. It is indicated on the pseudo code by bold font.

Algorithm 5 L1P1S1. Executes top-k spatial preference query utilizing simultaneous computation of many target objects' partial values, object discarding, and simultaneous computation of object's partial values based on similar preferences

```
L1P1S1 (tree root of index storing target objects R, preference set of the query P, parameter k)
1.  Create sets S_i containing preferences F_c∈P regarded as mutually similar.
2.  Create list I containing sets S_i ordered by their cardinality.
3.  Create pruning value x = 0.
4.  Create a min-heap D used for storing result objects ordered by their total value
5.  For every leaf L of R execute:
6.      Create set V containing target objects p stored at L.
8.      For every S_i ∈ I:
9.          If preferences from S_i use range neighborhood definition method:
10.             Compute partial values t_d(p) of objects p ∈ V executing GroupRNGValue(tree
                root of data specified by preferences from S_i, V, distance parameter of
                preferences from S_i, S_i).
11.         Otherwise, if preferences from S_i use nearest neighbor neighborhood definition
            method:
12.             Compute partial values t_d(p) of objects p ∈ V executing GroupNNValue(tree
                root of data specified by preferences from S_i, V, S_i).
10.         For every object p from V execute:
11.             Update t_+(p), i.e. the maximum possible value of object p, by subtracting from
                it the maximum possible value based on preferences from S_i set and then adding
                to it the computed partial value t_d(p)
12.                 If t_+(p) ≤ x:
13.                     Remove p from V. [object discarding]
14.     For every object p ∈ V:
15.         If the total value of p: t(p) > x:
16.             Insert p into D.
17.             If D contains more than k elements:
18.                 Remove first element in D.
19.                 Assign the value of the element currently first in D to x.
```

5. Algorithm Efficiency

In order to examine efficiency of algorithms utilizing different subsets of optimization techniques, queries presented in table 3 were being executed. The queries differ in terms of methods defining the neighbourhood and the number of similar preferences. Every object set contained 10,000 elements distributed uniformly in a two dimensional 1000 × 1000 constrained space. The queries were executed on Intel Pentium M 1.7GHz, 1.5GB RAM computer.

Table 3. Query character

query name	NN preferences.	RNG preferences	NN pref. similarities	RNG pref. similarities
8NN NoSim	8	0	0	N/A
8RNG NoSim	0	8	N/A	0
4NN/4RNG NoSim	4	4	0	0
8NN Sim	8	0	8	N/A
8RNG Sim	0	8	N/A	8
4NN/4RNG Sim	4	4	4	4
4NN/4RNG HSim	4	4	2	2
4NN/4RNG QSim/HSim	4	4	1	2

Figure 5 presents the number of index accesses performed by algorithms computing partial values of each target object individually. The data it presents indicates that queries containing many similar preferences require significantly less index accesses when executed by L0P0S1 algorithm utilizing the proposed optimization method, than when they are executed by L0P1S0 algorithm utilizing object discarding. For queries lacking similar preferences however L0P1S0 algorithm proved to be more efficient. In case of 4NN/4RNG QSim/HSim query which is composed of a moderate number of similar preferences, both algorithms require similar number of index accesses. Despite evident reduction of index accesses provided by optimization methods utilized by aforementioned algorithms, the choice of the one best fitted for the specific query is not straightforward. Therefore it has to be noted that it is possible to utilize both of these techniques simultaneously which is performed by algorithm L0P1S1. It executed all the queries at least as efficiently as the algorithm best fitted for that case. Moreover, in case of queries moderately prone to each optimization method utilizing both techniques simultaneously allowed further reduction in the number of index accesses compared to the algorithm best fitted for that case and utilizing one method.

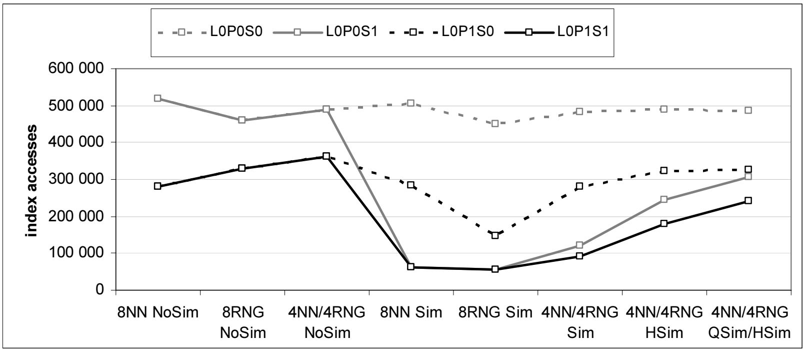


Figure 5. Number of index accesses realized using algorithms computing objects' values separately

Figure 6 presents the number of accesses to data realized during query execution using discussed algorithms. The results it presents are analogous to those of index accesses figure and, therefore, further increase the differences between algorithms' efficiency described earlier.

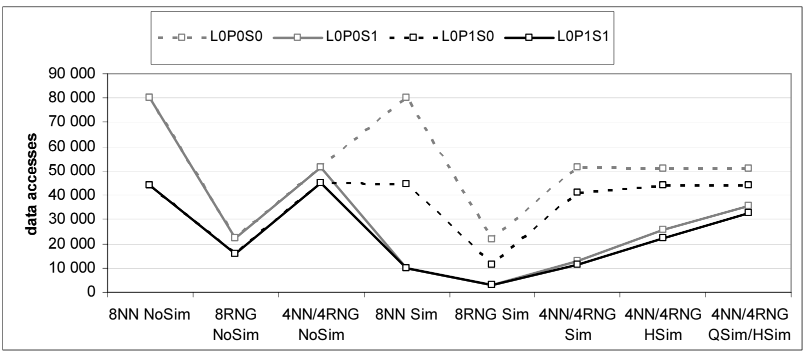


Figure 6. Number of data accesses realized using algorithms computing objects' values separately

Simultaneous computation of many target objects' values is another technique allowing reduction of the number of accesses. As this method often proves to be very efficient, it is imperative for the proposed optimization method to allow its use and not to cause its deterioration.

Figures 7 and 8 present respectively index and data accesses realized for different queries by algorithms computing many objects' values simultaneously. Both prove that the proposed optimization method integrates seamlessly with algorithms utilizing simultaneous computation of many objects' values. The optimization method's dependency on query definition discussed earlier is also valid for this class of algorithms. The algorithm utilizing all optimization techniques offer best performance regardless of query specifics. It was the most efficient algorithm in both separately discussed classes, as algorithms computing many objects' values simultaneously require significantly less accesses than algorithms computing each objects' value individually. This is indicated best by differences in graph scales.

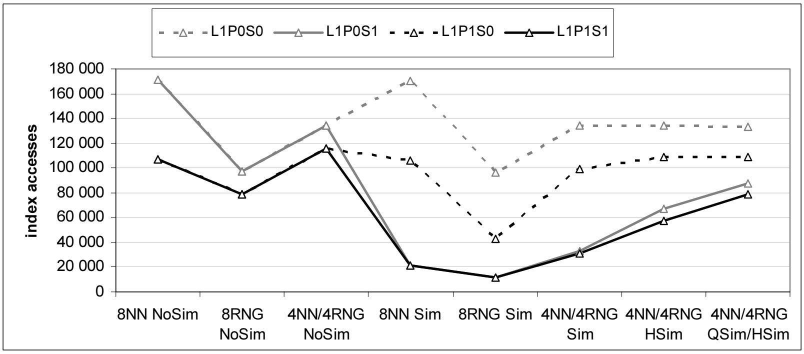


Figure 7. Number of index accesses realized using algorithms computing objects' values simultaneously

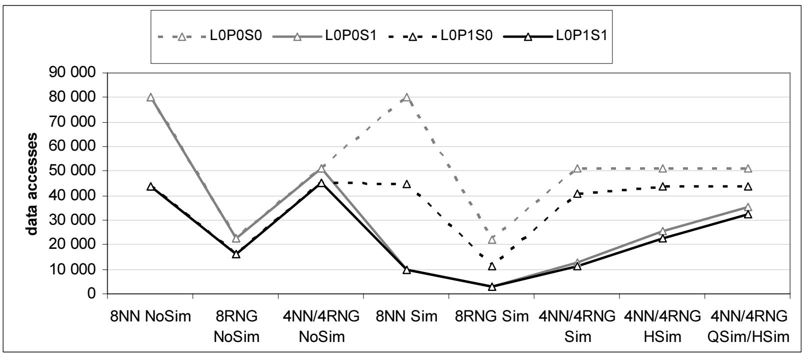


Figure 8. Number of data accesses realized using algorithms computing objects' values simultaneously

6. Summary

The proposed optimization technique reduces the number of index and data accesses for queries containing similar preferences. The reduction level is proportional to the number of similar preferences. For queries composed of many similar preferences this method is more efficient than objects discarding or simultaneous computation of their value. What is even more relevant is that when the introduced optimization is utilized along with existing techniques it can lead to significant increase in overall algorithm efficiency.

7. Further Work

At the moment, designing algorithms which would allow efficient execution of top-k spatial preference queries on distributed data is considered. This would allow us to execute the queries on data stored on remote computers and simultaneously to use their processing power.

References

[1] M.L. Yiu, X. Dai, N. Mamoulis, M. Vaitis: Top-k Spatial Preference Queries. *ICDE 2007*, 1076-1085.

[2] A. Guttman: R-Trees: A Dynamic Index Structure for Spatial Searching. *SIGMOD Conference 1984*, 47-57.

[3] Octavian Procopiuc. *Data Structures for Spatial Database Systems*.

[4] G.R. Hjaltason, H.Samet: Distance Browsing in Spatial Databases. *ACM Trans. Database Syst*. 24(2): 265-318(1999).

[5] N. Beckmann, H.-P. Kriegel, R. Schneider, B. Seeger: The R*-Tree: An Efficient and Robust Access Method for Points and Rectangles. *SIGMOD Conference* 1990, 322-331.

This page intentionally left blank

Information System Applications

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-155

Technical Metadata and Standards for Digitisation of Cultural Heritage in Poland

Grzegorz PŁOSZAJSKI
Warsaw University of Technology
Institute of Automatic Control and Computation Engineering & Main Library
Pl. Politechniki 1, 00-661 Warsaw, Poland
e-mail: ploszajski@bg.pw.edu.pl

Abstract. The digitisation of national heritage is supported in Poland by the government who want to establish standards and recommendations that could be used by all types of cultural institutions and organizations involved in digitisation process, including museums, libraries and archives. This paper describes an approach to establishing standards concerning technical aspects of digitisation, first of all technical and structural metadata and a set of rules, parameters and formats. The presented solutions are based on an unpublished report prepared by a working team.

Keywords. digitisation, technical metadata, cultural heritage, digital preservation

Introduction

Digitisation of cultural heritage consists in the creation of digital objects containing "images" or "copies" of physical source-objects together with appropriate information. Typical source objects are books and journals, archive documents, paintings, photographs, sculptures, analogue recordings of the sound, movies etc. Digital objects have a form of computer files, recorded usually as graphic images, text files, moving pictures, digital recordings of the sound etc. Digital objects contain also information on the digital object such as description of the digitised source object, process of digitisation, manner of presentation, rules of access, structure of the content of digital object and other data that can help management of digital objects. This information is called metadata ("data about data").

The primary aim of digitisation is ensuring a wide access to digital "copies" of objects belonging to the cultural heritage. The secondary aim is the preservation of these digital copies over time.

The digitisation of cultural heritage involves museums, libraries, archives and various other organizations. In Poland, the process of digitisation is supported by the government. The government intends to establish standards and recommendations that could be used by all types of cultural institutions and organizations involved in the digitisation process. Standards make the search, management and long term preservation of digital objects easier.

Poland is at a relatively early stage of the process of mass digitisation of cultural heritage, so we can learn from more experienced countries and use the solutions which they have applied successfully. New organizational solutions can also be considered, e.g. common repositories for long term preservation for archives, libraries, museums and other organizations involved in digitisation.

1. The endeavour to the standardisation

1.1. Typical digitised objects

Activities for providing access to cultural goods in the form of access to their digital copies and preservation of these copies have been undertaken in many countries on wide scale during the last twenty years. Variety of physical objects belonging to collections of libraries, archives and museums (books and periodicals, files of archival documents, plans and maps, photographs, paintings, technical drawings, sound records, video, movies, sculptures, arms, dishes, clothing, archaeological artefacts and many more) implied the variety of types of corresponding digital objects: simple text, still flat pictures, graphic 3D images, moving pictures, digital recordings of the sound or even combinations, like text linked to a graphic file (e.g. in case of manual rewriting of a handwritten document) or sound linked to graphic files.

Each kind of digital objects mentioned above can be recorded in many formats, with varying technical characteristics, with the legal conditions of usage (open formats or commercial) and with the ease of usage by the users (popularity). In addition, different technical parameters of digitisation can be applied, e.g. resolution, quantities of colors or sampling rates of the sound or the moving picture — having impact on the quality of digital object and its conformity or fidelity to the source object.

1.2. General guidelines

During the early period of digitisation many different solutions were developed and applied. As a result, the need for standardization became more clear and was expressed usually in form of guidelines. Guidelines used to be formulated for given projects, but when the libraries and the archives gained some experience, they were also formulated for usage in nationally. The guidelines recommended using some formats of computer files, values of technical parameters and formulated quality measures and procedures. Such guidelines are still in use. They help manage the policy of digitisation and preservation on the national scale or in an institution.

Digitisation guidelines have been formulated by some national libraries, e.g. in Australia [1] and New Zealand [2], and also by national archives e.g. in US [3]. Guidelines have been formulated also by UNESCO [4]. Initiatives funded by EU, such as Minerva, addressed their guidelines to all European countries [5]. Guidelines have also been formulated for local projects, e.g. “North Carolina ECHO” [6], or for institutions, e.g. Harvard University. Besides the digitisation guidelines there are also guidelines concerning digital archiving, e.g. prepared by the national archives of Australia [7]

The guidelines do not satisfy all needs. In cases where no guidelines can be found for a given problem, we can look for “best practices”, e.g. in similar institution or in a document like [8].

1.3. File formats

Some formats of computer files proved to be well suited for given digitisation tasks and became recommended and more often used. For example, specification Tiff 6.0 [9] for graphic files which was created by Aldus mainly for printing industry, has been widely accepted for making "master copies" of digitised objects.

Tiff is a lossless format. Another advantage of this specification is its compatibility with the industrial standard Exif [10] developed by the association of Japanese producers of photographic equipment JEITA. This standard allows to include metadata into graphic files. Accepted by many producers of the photographic and optical equipment, Exif is very useful for digitisation because computer programs can read technical metadata from graphic digital files.

Another example of a graphic file capable of containing metadata in Exif format is JPEG [11], created by Joint Photographic Experts Group. It was also designed to be an interchange format [12]. JPEG became popular as a format for network access of digitised objects because due to lossy compression, files are much smaller than equivalent tiff files and JPEG has been widely accepted by web browsers (however, the reason of its popularity was due to its wide use in popular digital cameras rather than to digitisation).

Many other standards were designed so as to be capable of including metadata. This seemed to be important in case of sending digital objects to repositories, because all information concerning one object was kept together. An example of such standard is the IPTC4XMP [13], prepared by International Press Telecommunications Council (IPTC) together with the firm Adobe and IDEAlliance (International Digital Enterprise Alliance); the standard makes it possible to include descriptive metadata into TIFF and JPG files. Thanks to such solution, a graphic file sent to the agency by a photographer in the form of TIFF files already has a recorded complete set of metadata.

There are also formats of audio files and video files which are capable of containing metadata, e.g. [14].

Sustainability is another requirement concerning formats when long term preservation is required [15], [16] and [17].

There are other formats of graphic files used for digitisation in some countries, e.g. JPEG2000, DjVu and MrSid. However, in Poland TIFF and JPEG have a privileged position in projects which received public funds.

1.4. Technical Standards

Recommendations concerning digitisation activities within single projects became a foundation to build more general standards. Most efforts were focused on metadata and there are not many formal standards related to digitisation in general and to its technical parameters. The following documents are called "Technical standards", however, rather in a common sense than formally (none of them is approved by national or international standards office).

"The Library of Congress Technical Standards for Digital Conversion of Text and Graphic Materials" [18] designed for the "American Memory" project is an example of technical standards. Another example is "Technical Standards and Guidelines for CCO (Canadian Culture Online) Funded Initiatives" [19].

A detailed information on technical standards used in some national digitisation programs in the Czech Republic is available in [20].

"Summary of LC Image Quality Standards by Document Type" is an example of quality standards (used in "American Memory" project) [21].

The above mentioned standards concern digitisation of graphic materials. Technical standards are formulated also for digitisation of audio and video materials, e.g. Australian AIATSIS Audiovisual Archive [22], a very simple 2-page document, based mainly on UNESCO guidelines.

1.5. Metadata Standards

The metadata concerning digitised objects are divided into various groups: descriptive, technical, administrative, legal, behavioral, structural and preservation metadata:

- descriptive metadata describes the digitised source-object (creator of the source-object, date of the creation, type of object, title etc.);
- technical metadata concerns the process of creation of a digital image (name and version of used software, format of files, size of files, the quality of digital image expressed with the resolution or with the sampling rate etc.);
- administrative metadata is another data that can help manage the digital objects (creator of the digital object, date of creation, type of digital object etc.)
- legal metadata concerns rules of access, copyrights to the object etc.;
- behavioural metadata concerns recommended manner of presentation of the object (e.g. vertically or horizontally);
- structural metadata concerns the structure of the content of the digital object (e.g. partition of the book not only on pages, but also on chapters, partition of the broadcast, synchronization of images with the text and audio in a multimedia presentation etc.);
- preservation metadata describes all changes of a digital object during its life cycle (e.g. migration to new formats or changes in metadata).

In a less detailed version, administrative metadata contain also some other groups, e.g. technical or legal.

There are many publications concerning the metadata standards used for digitisation purposes, e.g. [23] made by Technical Advisory Service for Images in UK.

1.5.1. Descriptive metadata

The descriptive metadata standards are not in the scope of this article. There is a well known Dublin Core standard [24] (ISO Standard 15836), worked out to facilitate describing of e-documents by persons who do not have the librariansŠ skills in cataloguing. The DC standard covers also legal metadata concerning the access rights. It has limited possibilities of presentation of the technical metadata.

If the Dublin Core could be approved by the libraries, archives and museums as a standard of description metadata for digitisation — or rather for wide (public) access to digitised objects — then users could easily and simultaneously search information in all these types of cultural institutions. The main problem is the relation of standards (formats) used for description of source objects in libraries, archives and museums to the DC standard (or any common standard). The libraries are in the best situation because the format MARC 21 [25], widely used in libraries, can easily be transformed to the Dublin Core. Archives have different approach to their holdings from the libraries; they

often describe files rather than individual documents. Archives often use the EAD [26] standard (format) of such description. Conversion of such information from EAD to DC is problematic.

Holdings of museums have a more differentiated character than those of archives and libraries. There is no one dominating metadata standard for museums. Reviews concerning metadata standards can be found in works of ICOM-CIDOC (The International Committee for Museum Documentation of the International Council of Museums), such as [27] and [28]. Canadian Heritage Information Network (CHIN) [29] is an example of a concrete approach to standards and guidelines with detailed instructions on how to deal with various types of museum objects.

1.5.2. Technical metadata and preservation metadata

Among standards prepared by the Library of Congress [25] are: MIX (Metadata for Images in XML), PREMIS (Preservation Metadata) and TextMD (Technical Metadata for Text). MIX and TextMD are designed to encode technical metadata for digital raster still images and text based digital objects. PREMIS defines core metadata for long term preservation. All three standards have a form of XML Schema. For MIX and PREMIS also data dictionaries are defined. The data dictionary for MIX 1.0 became an American standard ANSI/NISO Z39.87, approved in 2006 (not fully compatible because current version of MIX changed in 2008 from 1.0 to 2.0); the one for PREMIS was released in April 2008.

The Library of Congress was also working on metadata standards for audio and video materials [30]; available documents have historical character to 2004. There is no information about continuing these works since 2003.

Metadata standards used for images by the National Library of Australia [1] are based on older version of Z39.87 standard (Working draft, 2000).

The National Library of New Zealand having learned from experience of Australian and American libraries, introduced own modifications and formed a standard which comprised technical metadata and preservation metadata not only for traditional library materials, but also for the audio, video and text materials, as well as computer data files and system files. At the same time technical recommendations have also been formulated. The metadata are listed in "Preservation Metadata. Metadata Standards Framework — Metadata implementation schema" [31]. They are divided into four groups concerning: 1. the digital object (19 metadata), the process of creation of this object (13), the resultant file (11 general elements plus additionally 8 for the image, 6 for the audio, 8 for the video and 2 for the text), and the information on changes made in metadata during the life cycle of digital object (5 elements). Many technical metadata were defined as equivalent to some metadata from current version of Z39.87 at the time of creating the standard (draft version for trial use, 2002). In the 2008 they have still equivalent metadata in the ANSI/NISO Z39.87 approved in 2006 (only ID numbers of metadata have changed). Metadata for the audio files were compatible with the standards of the European Broadcasting Union.

Some countries which decided to introduce own standards followed standards already designed. Standard of preservation metadata LMER introduced in Germany by National Library was inspired by the New Zealand approach. Standard of administrative and technical metadata introduced in Italy was based on Z39.87 (draft version).

1.5.3. Structural metadata

Among standards able to express the structure of complex digital object two seem to be the most important for digitisation: METS and MPEG-21 DIDL.

METS (Metadata Encoding and Transmission Standard), designed and maintained by the Library of Congress [25], was created to describe the structure of complex objects of digital libraries. Standard has additional ability to contain descriptive, administrative and other groups of metadata, as well as names and locations of files constituting that digital object. Thereby METS can be used to the transfer of complete information on digital objects. However it is only the structural part of metadata that is obligatory in METS file. The METS standard is described in details in "METS. Primer and Reference Manual" [32].

Standard MPEG-21, designed by Moving Pictures Experts Group, is called a multimedia framework. Some parts of this standard are approved as ISO/IEC 21000 standard. Among them is the second part, publicly available, concerning the Digital Item Declaration (DID) and Digital Item Declaration Language (DIDL) [33].

Both standards: METS and MPEG-21 DIDL have nearing targets, though MPEG-21 is conceived and practical mostly for the purpose of the "wrapping" of audio-visual materials [34], [35], while METS more often to "packing" of collections of still pictures.

These two standards are flexible enough to express many various structural relations. The third standard — Florida State University Metadata Standard [36] — is not a structural metadata standard. It is a complete set of metadata for current use in digital library: administrative, descriptive, technical and structural (without preservation metadata). The main feature of this standard is its simplicity in comparison to the Library of Congress standards. The FSU standard is oriented to typical library collections, so in expressing complex structure is less flexible.

The structural group of FSU metadata is very simple. There are only four structural metadata, each with several options. These options have direct relation to typical library objects. All metadata are defined in FSUMD data dictionary.

The FSU metadata standard represents quite different approach than METS and MPEG-21. It even is used with METS (as transportation standard).

1.6. Technical requirements for digitising

1.6.1. Still images

In the document "Digitisation guidelines. Specification for Imaging"[2] are included recommendations for digital master files and for derivative files (screen images and thumbnails). Recommendations for master files specify file format, capture resolution, bit depth and colour space. For derivative files was specified one version of recommendations.

Similar recommendations for master files were formulated by the Library of Congress for "American Memory" project [18]. They specified the resolution, bit depth, greyscale factors and colour accuracy. Greyscale was characterized by number of distinguished grey levels on given (20 levels) target and by noise. The colour accuracy was expressed by Delta E factor and ICC profile. For each category of digitised materials two versions of "expected outcome" were specified, characterizing the aim of digitisation, e.g. access to content, recognition of artefactual features, reproduction or research.

Requirements formulated in this paper are inspired mainly by these two documents.

1.6.2. Audio, video, moving images

Some recommendations concerning the audio and video files have been formulated in the mentioned Australian AIATSIS Audiovisual Archive [22]. Information on digitising the sound recording can be found in the Library of Congress document "Audiovisual Prototyping Project" [37], which is an abbreviated and modified version of a document drafted in 1999. A more detailed information concerning digital moving images and sound archiving can be found in a study [38]. A lot of recommendations and opinions is expressed in a report of a roundtable discussion of best practices for transferring analog discs and tapes [39]. Some recommendations can be found in publications of International Association of Sound and Audiovisual Archives, e.g. on the safeguarding of the audio heritage [41].

1.7. Special types of objects

1.7.1. Objects 3D

There are several standards of 3D objects, among them open standards as VRML or X3D, however there are neither technical metadata standards for digitising such objects and for long term preservation.

1.7.2. Text and OCR

The textual data can be introduced by hand, but on small scale. The basic meaning in the mass-digitisation has an automatic extraction of text from graphic images of pages with printed or handwritten text by means of OCR programs.

There is a problem of defining relation between the recognized text and the graphic image of this text. Such problem is solved practically e.g. in pdf files. In case of digitisation an ALTO standard [40] can be used; this standard relates each recognized word to corresponding fragment of graphic image. Standard ALTO was made and is maintained by CCS Gmbh, however it can be used as open standard. It is often used in case of mass digitisation of newspapers. It has a form of the METS profile, so it can easily be used with METS.

Another problem concerns the quality of text recognition and the use of recognised text by the users. Usually some errors happen. If the text is to be shown to the users, the error rate should be small. If it is not small, then the users can be shown only graphic image of page, while they can use text search of the document. If the search engine finds a word, then corresponding piece of the graphic image can be marked off. However the user can not be sure whether he managed to find all appearances of searched text in the document.

2. Initial recommendation of standards

2.1. Situation in Poland

The digitisation of the national heritage in Poland was being made on a rather small scale during many years. The libraries were perhaps more active than archives and museums, however their efforts were addressed rather to giving access do digitised books and doc-

uments then to long term preservation. Lossy formats of graphic files were often applied, e.g. DjVu in a group of digital libraries (dLibra), because of faster transfer of images through computer network and smaller (thus cheaper) requirements for disk space or other storage.

During this period institutions engaged in digitisation gained some experience, even if some activities done in this field might be called chaotic. Now Poland prepares to the mass digitisation. There are three main problems that should be addressed:

1. establish official standards which would help to do the digitisation in good order,
2. organize repositories for long term preservation for digital objects,
3. prepare guidelines and organize training centres, which would instruct and help smaller or less experienced institutions how do proceed with digitisation.

These standards would help not only institutions doing the digitisation of national heritage but also the government agencies which could formulate digitisation policy and specific requirements in a more clear way.

2.2. To build standards or to accept known ones

The decision to accept chosen existing standard has many advantages, especially if the standard had been used for some time and proved correct. Some software tools compatible with such standard can be available, e.g. for the extraction of metadata from digital images. However such decision brings a problem of the following of future changes and modifications of the standard. Another aspect is the need of training to understand technical metadata, because some standards, e.g. MIX (Z39.97), use a lot of advanced knowledge concerning the optics and the digital imaging.

An own standard has also some advantages. The users do not depend on changes in external standard. The software able to control metadata by means of XML Schema can be prepared easily. If some level of interoperability with external standards should be achieved, then data from own standard could be converted to external standard. This kind of interoperability might be assured rather by repositories then individual institutions.

The standards prepared by the Library of Congress are being treated by many like a model solutions. They can take pattern from them while building own solutions or try to achieve some level of compatibility or at least compare their standards to those of the Library of Congress. It concerns especially standards of technical metadata.

If we asked about "good" standard for the technical metadata considering its level of excellence, then perhaps MIX would be the answer in case of the still 2D raster images. Similarly if we asked about good standard of structural metadata, METS could be mentioned.

However the choice of standard must take into account current level of digitisation skills of cultural institutions in a coutry. Both standards are rather complicated. If everything or almost everything was done automatically by convenient software then these standard might be used widely. But such solution is costly. If the metadata were prepared partly manually — what is probable in small institutions — then complexity of standard might cause additional level of difficulty.

2.3. Two groups of standards

In such situation two groups of standards can be considered:

1. simple temporary standards to be implemented in the near future,
2. mature standards of destination, which can be introduced later.

Introducing a metadata standard before long term repositories are established (such repositories do not exist yet) puts duties to the repositories which will be obliged to accept all digital objects with metadata conforming to the standard.

Another consequence of introducing a metadata standard concerns the institutions digitising the national heritage. They should analyse whether information gathered in their digital libraries and repositories is complete and could be converted to such standard and exported. They could revise their procedures with respect to gathering technical metadata and conventions of giving names to the files. These institutions do not have to use these metadata standards directly in their local systems — neither now nor in the future — but they must check their ability to export this information in the form conforming to the standard (or one of accepted standards).

Temporary standards could be verified whether conversion to any destination standard is possible (it is an advantage, not a must. Possible data structure in which repository would keep information from both standards could also be analysed.

The destination standards do not have to wait long for acceptance if there is no doubt that they should be accepted in the future. Such standards could be used by experienced institutions if they did not want to use temporary standards or if they need to use such attributes of these mature standards which are not available in the temporary ones (e.g. complex structure).

2.3.1. Destination standards

Among the mature metadata standards are MIX and METS.

The proposition is to accept METS as the standard able do express complex structural relations. Next proposition is to accept ALTO as standard of expressing relation between the text recognized by means of OCR software and the graphic image of each page. ALTO works well with METS, what makes another reason for accepting both.

The MIX (Z39.87) standard is proposed to be accepted or at least reconsidered in the near future (e.g. two years) after having gained more experience. It should also be considered whether to choose MIX 1.0 or MIX 2.0 or another future version of this standard. Anyway, temporary standard should be able to convert its data (or part of its data) at least to the MIX 1.0 (Z39.87).

The TextMD standard of technical metadata for text based objects is a new one. The work on this standard is in progress. It is defined as an extension schema for METS, however can be applied also without METS. It looks interesting and should be considered in the future as a candidate for destination standard. It could also be included in the temporary standard as an option.

It seems that in case of technical metadata for digitising audiovisual files the standardization process needs some time to prepare mature, widely accepted metadata standards. Digitising institutions can make decisions depending on the state of objects and risk of deterioration of analogue carrier and either wait or digitise using some chosen temporary standards (or guidelines).

2.3.2. ABMPL — simple temporary standard

A temporary metadata standard called ABMPL has been designed. Its acronym stands for Archives, Libraries (Biblioteki in Polish), Museums and PL for Poland. This standard

has already been designed and is ready to be presented soon to Polish community of libraries, archives and museums. This standard is presented in the next section.

2.4. *Temporary revision of standards*

It is assumed that metadata standards should be revised at least every three years. The standards can be modified. New standards can be approved and used in parallel with the older ones.

If recommendation for given standard expired it should be interpreted so that during some defined period main repositories would be obliged to accept digital objects sent to them with metadata in such standard. Only new projects should not use this metadata standard. Repositories generally convert such metadata to their internal data structure. They might also convert these metadata to a new standard or prepare a software tool that might do such operation for institutions doing the digitisation.

3. Proposed metadata standard for archives, libraries and museums in Poland

The proposed metadata standard ABMPL intends to be a complete structure, having containers for various groups of metadata, among them also for description metadata. The structure of metadata groups is inspired by the New Zealand metadata standard [2], called NZ, however the structure of group "file" is inspired mainly by the FSU [36] standard.

3.1. *Groups of metadata*

There are seven groups of metadata:

1. Object — based on Object group of NZ standard with changed identifier section: 12 metadata plus repeatable container of two metadata for identifiers (solution inspired by MIX and PREMIS). Among obligatory identifiers are local identifier, parent local identifier and universal prefix changing local identifiers to URN or to PURL, depending on decision of Polish authorities.
2. Process of creation and conservation of digital object — 12 metadata from Process group of NZ standard plus a container for check sum fixity (followed by MIX).
3. Technical metadata — a container with subgroups and files ("subgroup" and "file" is a container and name of metadata having own attributes — after FSU). Each file has a header group of 9 metadata (coming from File group of NZ standard — with minor changes) and Byte Order metadata from MIX. The header group is followed by one (or two in case of video and audio) of the following six groups containing technical metadata for objects of type defined in the MIME-type metadata:

   (a) Image — this group has four subgroups:

      i. 12 metadata from Image group of NZ standard,
      ii. optional container for 25 photo metadata (Exif standard) used in MIX in subgroup Camera Capture Settings,

   iii. optional container for 31 GPS metadata (Exif standard) used in MIX,
   iv. 4 metadata — coordinates of main part of the master image without the targets (optional).

  (b) Audio — 7 metadata from Audio group of NZ,
  (c) Video — 10 metadata from Video group of NZ,
  (d) Text — 2 metadata (character set and markup language) from Text group of NZ,
  (e) Data Set — 1 metadata (after NZ),
  (f) System File — e.g. needed to use digital object (1 metadata — after NZ).

4. Metadata Modification — a repeatable container with 4 metadata taken from similar group in NZ (preservation metadata).
5. Descriptive Metadata — a container for descriptive, legal and other metadata conforming to Dublin Core standard (allowing for common search among library, archive and museum objects).
6. Additional Descriptive Metadata — two containers for descriptive metadata from original standards for future use (e.g. a better conversion from original standard to Dublin Core or a new methods of search); one container for XML metadata, one for non XML (encoded in Base64).
7. Administrative and Legal Metadata — 7 metadata from FSU.

3.2. Structural metadata

Structure of a complex object can be expressed with attributes of two metadata: subgroup, file, taken from FSU [36]. Subgroup is a repeatable container for the files; it has four attributes: type, id, sequence, head (id is obligatory). The attribute "type" is to be chosen from predefined list, containing: collection, volume, issue, chapter, page, correspondence and others; attribute "head" contains text shown to the user; "sequence" defines the order of presenting subgroups to the user. Original list of types was enlarged with "linked" to link e.g. text and graphics.

Attributes of the file are: type and head. List of types of the file is similar to that of subgroups (minor differences). Files are presented to the user in order they have in a subgroup (e.g. pages of a book).

3.3. Two types of metadata files

It is proposed that there will be two types of files with metadata:

1. technical-preservation (mTP),
2. structural-descriptive (mSD).

Each computer file being the digital "copy" of analogue object or its part has one metadata file with technical-preservation metadata (mTP). Name of computer file is included in the name of this metadata file.

Each digital object being a "copy" of digitised analogue object has structural-descriptive metadata file (mSD). This group of metadata files can (should) be named in different manner to that of technical-preservation metadata.

Thus a single page document will have one mTA file and one mSD file while a 50 pages book will have 50 mTA files and one mSD file.

Such structure can ease transactions of sending files to repositories and making conservatory operations during temporary period (before creation of main repositories).

4. Proposed recommendations for digital still images of flat motionless objects

4.1. Introduction

Recommendations of the New Zealand National Library concerning the "imaging" [2] formulate two levels of requirements: "minimal" and "recommended"; the goals of digitisation are not considered. Instead recommendations of the Library of Congress formulate the aim of digitisation for each category of physical objects, e.g. in case of images of text documents there are two options: either to give the user view of the content or to recognize the text by means of OCR and make it searchable.

In the presented solution an intermediary approach has been chosen. Two levels of technical requirements are defined: the "minimal" and "recommended" (like in [2]), but several general rules have been added which should be taken into account to interprete correctly the two level of requirements. These rules are addressed also to non experienced institutions and organizations which could suffer lack of training; that is why some of the rules have form of basic reminders.

4.2. Proposed general rules — in brief

The rule 1 — suitable and correctly calibrated "capturing" device is a principal condition of making good digital copies (the "suitable" is not equivalent to "expensive").

The rule 2 — correctly calibrated monitor is a necessary condition of the man made optical qualitative control within the grey range and the colour range. To the calibration of the monitor one ought to use charts (targets) recommended by the producer.

The rule 3 — black and white one-bitwise images are subject to the optical qualitative estimation which consists in the check whether elements of original object are presented correctly, without the deformation of graphic characters, or whether did not come into being dark "stains" misinterpreting the original.

The rule 4 — the greyscale is based on no fewer than 8 bits. Parameters of the digital image are subject to the optical qualitative estimation which consists in: the check whether the grey target was mapped correctly and whether details in parts of tones very clear or very dark are correctly presented. The optical man made estimation can be supplemented with the measurement made automatically (gamma, noise, dynamics).

The rule 5 — the colour should be used in images instead of the greyscale, when it is an essential attribute of the document. Should this be then no fewer than 24-bitwise. The colour in the digital picture image is subject to the optical estimation which consists in: the check whether the colour target was mapped correctly and whether details in parts of tones very clear or very dark are correctly presented. The man made optical estimation can be supplemented with the measurement made automatically (concordance of colours with the target, noise, dynamics).

The rule 6 — printing of colour reproductions needs additional information on the ICC profile. As the minimum-requirement the profile ICC AdobeRGB 1998 is recommended.

The rule 7 — the resolution of images of documents and large physical objects: if for a given type of object the recommended resolution of digital image is expressed both as a number of pixels per inch (ppi) and as a number of pixels per its greater dimension, then the smaller number of pixels per object should be applied. In the case of documents without petty details (e.g. school wall-sketch-maps, most of posters and bills, photographic prints of very large sizes) the resolution may be smaller then recommended.

The rule 8 — the resolution of images of small objects and/or documents with petty details: if recommended resolution is expressed both as a number of pixels per inch (ppi) and as a number of pixels per its greater dimension, then the greater number of pixels per object should be applied. The same refers to museum objects except the large ones. For documents printed in very small type the resolution should be even greater from recommended (e.g. inversely proportional to the size of types, assuming that standard recommendations are addressed to 10-12 points type); similarly in case of small photographic prints.

The rule 9 — grey or colour and grey targets should be placed near the digitised object (in a distance that makes possible their separation on the digital image). In case of objects with repeatable parameters, e.g. following pages of a book, images with the targets can be made separately for representative number of pages but not for the whole book.

The rule 10 — rational utilization of resources. The recommended requirements stated below can be enlarged in compliance with the rule No. 8. However the usage of this rule makes sense only then, when better representation of details has an important reason.

4.3. Recommended and minimal technical parameters of digitisation

4.3.1. Groups of objects

The recommended technical parameters were determined for seven groups of digitised materials (in NZ guidelines there were four groups). These seven groups are: 1. printed text with good contrast; 2. printed text with halftone illustrations or poor contrast; 3. monochrome pictures, graphics and photo prints; 4. negatives, slides; 5. microfilms; 6. paintings, colour photographic prints (also sepia etc.), colour drawings, copperplates etc., manuscripts, incunabula, and small and medium size museum objects and artefacts; 7. posters, big maps, big museum objects.

4.3.2. Minimum requirements

Recommended graphic file format for all seven groups is Tiff 6.0. Group 1 has 1 bit per pixel while others have either 8 bit (grey) or 24 bit (colour). Compression LZW is allowed for groups 2-6 but not recommended. For group 1 compression ITU G4 (CCITT fax Group 4) can be used. Minimum resolution for group 1 (1-bit) is 400 ppi, for groups 2 and 7 is 300 ppi, while for groups 3,4 and 6 is 300 ppi but not less then 3000 pixels on longer dimension. For group 5 (microfilms) it is recommended to use resolution such

as for the microfilmed original but limited to the resolution of microfilm. Grey targets should be used for groups 2 and 3 (Gamma 2.2). For groups 4, 6 and 7 ICC profile AdobeRGB 1998 is recommended. Groups 1 and 5 do not have such requirements.

4.3.3. Recommended requirements

Recommended graphic file format for all seven groups is Tiff 6.0. Group 1 has 1 bit per pixel, while 2, 3, 4, 6 and 7 have 16 bit (grey) or 48 bit (colour); group 5 (microfilms) has 8 bits per pixel for grey and 24 RGB for colour microfilms when optical density D is ≤ 2. Compression LZW is allowed for groups 2-6 but not recommended. For group 1 compression ITU G4 (CCITT fax Group 4) can be used. Minimum resolution for group 1 (1-bit) is 600 ppi, for group 2 is 400, for group 7 is 300 ppi, for groups 3 and 6 is 400 ppi but not less then 5000 pixels on longer dimension and for group 4 is 600 ppi but no less then 5000 pixels. For group 5 (microfilms) it is recommended to use resolution such as for the microfilmed original but limited to the resolution of microfilm. Grey targets should be used for groups 2 and 3 (Gamma 2.2). For groups 4, 6 and 7 ICC profile AdobeRGB 1998 or better is recommended. Groups 1 and 5 do not have such requirements.

5. Proposed recommendations for audio, video and film objects

It is generally recommended for the libraries, archives (document archives) and museums to outsource the digitisation of audiovisual materials to experienced and well equipped centres. It is also recommended to wait for the results of projects concerning the radio and TV archiving, which are in progress. If digitisation is urgent because the analogue carrier is in poor condition then institutions can rely on a set of “safe” temporary recommendations, similar to those of AIATSIS [22]. The audio recommendations differentiate quality materials for which higher parameters are suggested, e.g. 96 kHz /24 bit instead of 48 kHz / 24 bit in case of a record of artistic performance; a lossless audio compression can be applied. In case of video and film (moving pictures) materials recommendations depend on the technology of original materials and on the aim of digitisation. For archiving purposes the minimum bit-rate 25 MBit/s is recommended.

6. Final remarks

The EU Commission recommendation on the digitisation and online accessibility of cultural material and digital preservation [42] — being a part of i2010 strategy which aims i.a. at building a common European Digital Library [43] — obliges member states to: “ensuring that cultural institutions, and where relevant private companies, apply common digitisation standards in order to achieve interoperability of the digitised material at European level. . . ”. The metadata standard and technical recommendations presented in par. 3-5 are proposed to be officially accepted as a (partial) fulfillment of this obligation. The metadata standard can also play some role with respect to the long-term preservation of the digital material.

References

[1] National Library of Australia, *Digitisation of Traditional Format Library Materials*, rev. 2006, www.nla.gov.au/digital/standards.html

[2] Alexander Turnbull Library National Library of New Zealand, *Digitisation Guidelines. Specification for Imaging*, October 2006, www.natlib.govt.nz/catalogues/library-documents/digitisation-guidelines

[3] U.S. National Archives and Records Administration, *Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files — Raster Images*, www.archives.gov/research/arc/digitizing-archival-materials.html

[4] UNESCO, *Guidelines for the Preservation of Digital Heritage*, 2003, unesdoc.unesco.org/images/0013/001300/130071e.pdf

[5] Minerva, *Technical Guidelines for Digital Cultural Content Creation Programmes*, 2004, www.minervaeurope.org/publications/technicalguidelines.htm

[6] *North Carolina ECHO Digitization Guidelines*, http://www.ncecho.org/guidelines.asp

[7] National Archives of Australia, *Digital Recordkeeping: Guidelines for Creating, Managing and Preserving Digital Records*, 2004, www.naa.gov.au/records-management/publications/Digital-recordkeeping-guidelines.aspx

[8] *Western States Digital Imaging Best Practices*, 2003, www.bcr.org/cdp/best/digital-imaging-bp.pdf

[9] Adobe Developers Association, *TIFF Revision 6.0*, 1992, partners.adobe.com/public/developer/en/tiff/TIFF6.pdf

[10] Standard of Japan Electronics and Information Technology Industries Association *Exchangeable image file format for digital still cameras: Exif Version 2.2*, 2002, www.exif.org/Exif2-2.PDF

[11] International Telecommunication Union, *Recommendation T.81: Digital Compression And Coding Of Continuous-Tone Still Images — Requirements And Guidelines* , 1992, www.w3.org/Graphics/JPEG/itu-t81.pdf

[12] *JPEG File Interchange Format — Version 1.02*, 1992, www.w3.org/Graphics/JPEG/jfif3.pdf

[13] *IPTC Core*, www.iptc.org/IPTC4XMP

[14] European Broadcasting Union, *Document Tech 3293: EBU core Metadata set for Radio archives*, 2001

[15] *Sustainability of Digital Formats*, www.digitalpreservation.gov/formats

[16] *Sustainability of Digital Formats Planning for Library of Congress Collections*, www.digitalpreservation.gov/formats/content/still_preferences.shtml

[17] *File Formats for Preservation* www.erpanet.org/events/2004/vienna/index.php

[18] *The Library of Congress Technical Standards for Digital Conversion of Text and Graphic Materials* , 2006, memory.loc.gov/ammem/about/techStandards.pdf

[19] *Technical Standards and Guidelines for CCO(Canadian Culture Online) Funded Initiatives*, 2004, www.pch.gc.ca/progs/pcce-ccop/pubs/techGuide_e.cfm

[20] Memoria Digitization, *Technical Standards Concerning Digitization* digit.nkp.cz/techstandards.html

[21] *Summary of LC Image Quality Standards by Document Type*, 2006, memory.loc.gov/ammem/about/standardsTable1.pdf

[22] AIATSIS Audiovisual Archive, *Technical Standards and Guidelines for Digitisation of Materials*, 2007, www.aiatsis.gov.au/__data/assets/pdf_file/7034/Technical_Standards_inc_video_V8.pdf

[23] *Metadata Standards and Interoperability. Technical Advisory Service for Images*, 2006, www.tasi.ac.uk/advice/delivering/metadata-standards.html

[24] *Dublin Core Metadata Initiative Element Set (ISO Standard 15836)* dublincore.org/documents/dces

[25] *Standards at the Library of Congress* www.loc.gov/standards

[26] *EAD (Encoded Archival Description)* www.loc.gov/ead

[27] *Museum information standards. ICOM-CIDOC*, www.willpowerinfo.myby.co.uk/cidoc/stand0.htm

[28] *Building semantic bridges between museums, libraries and archives: The CIDOC Conceptual Reference Model*, www.firstmonday.org/issues/isuue9_5/gill/index.htm

[29] *Metadata Standards for Museum Cataloguing* www.chin.gc.ca/English/Standards/metadata_intro.html

[30] *AV Prototype Project Working Documents: Data Dictionary for Administrative Metadata for Audio, Image, Text and Video Content* www.loc.gov/rr/mopic/avprot/extension2.html

[31] National Library of New Zealand, *Preservation Metadata. Metadata Standards Framework — Metadata Implementation Schema*, July 2003, www.natlib.govt.nz/catalogues/library-documents/downloadpage.2007-02-15. 6613783926

[32] *METS Primer and Reference Manual*, 2007, www.loc.gov/standards/mets

[33] ISO/IEC 21000-2, Second edition, 2005, *Information technology — Multimedia framework (MPEG-21) — Part 2: Digital Item Declaration*, standards.iso.org/ittf/PubliclyAvailableStandards/c041112_ISO_IEC_21000-2_2005(E).zip

[34] J. Bekaert, P. Hochstenbach, H. Van de Sompel, *Using MPEG-21 DIDL, to Represent Complex Digital Objects in the Los Alamos National Digital Repository*. D-Lib Magazine November 2003, www.dlib.org/dlib/november03/bekaert/11bekaert.html

[35] J. Bekaert, X. Liu, and H. Van de Sompel, *Representing Digital Assets for Long-Term Preservation using MPEG-21 DID*,DCC Symposium: Ensuring Long-term Preservation and Adding Value to Scientific and Technical Data (PV 2005), Nov. 21-23, 2005, Edinburgh, library.lanl.gov/cgi-bin/getfile?LA-UR-05-6878.pdf

[36] *Florida State University Metadata Standard. An Introduction to FSUMD* new.lib.fsu.edu/dlmc/dlc/fsumd_intro

[37] The Library of Congress, *Audio-Visual Prototyping Project*, 1999, www.loc.gov/rr/mopic/avprot/audioSOW.html

[38] Arts and Humanities Data Service, *Moving images and sound archiving final report*, 2006, ahds.ac.uk/about/projects/archiving-studies/index.htm

[39] Council on Library and Information Resources, *Capturing Analog Sound for Digital Preservation: Report of a Roundtable Discussion of Best Practices for Transferring Analog Discs and Tapes*, 2006, www.clir.org/pubs/abstract/pub137abst.html

[40] *METS / ALTO XML Object Model* www.ccs-gmbh.com/alto/

[41] International Association of Sound and Audiovisual Archives, *IASA-TC 03: The Safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy*, 2005, www.iasa-web.org/downloads/publications/TC03_English.pdf

[42] EU Commission, *Recommendation on the digitisation and online accessibility of cultural material and digital preservation*, 24 August 2006, ec.europa.eu/information_society/activities/digital_libraries/commission_recommendation/index_en.htm

[43] *i2010: Digital Libraries Initiative* ec.europa.eu/information_society/activities/digital_libraries/index_en.htm

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-171

Translation Accuracy Influence on the Retrieval Results

Jolanta MIZERA-PIETRASZKO and Aleksander ZGRZYWA
Institute of Applied Informatics, Wrocław University of Technology
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
e-mail: jolanta.mizera-pietraszko@pwr.wroc.pl, aleksander..zgrzywa@pwr.wroc.pl

Abstract. This report presents a novel methodology in evaluating performance of the popular on-line multilingual search engine AltaVista. We study some crucial aspects of natural language that usually disrupt translation process and the extend to which it influences retrieval results. Having prepared the test set, we analyze phenomena of an English and French language pair in relation to the strategy of browsing the Web for documents with the features specified in the query. Using Natural Language Processing techniques, we test a Machine Translation system performance in order to improve the translation quality, which in turn has an impact on the results of information retrieval in a language other than the query language.

Keywords. natural language processing, cross-language information retrieval, machine translation

Introduction

Performance of retrieval over languages aimed at increasing accessibility to on-line documents is a core objective of the research projects in CLIR (Cross-Language Information Retrieval).

In the process of multi-language information retrieval, in which the system retrieves the documents in all languages that they occur in the databases searched, and cross-language IR (the system translates the query to search for the documents in a language indicated by the user), the key strategies focus on both translation quality of the user's query as well as the adequacy of the document content to the user's need accordingly.

These two factors influence the overall system performance to almost the same degree. Any mismatch at the stage of the source language translation is a serious impediment to the number of documents retrieved. Simultaneously, any ambiguity occurring in the query affects the system performance as well. Fuzzy idea about the user's need is the other decisive factor for inadequate system responses [3].

In general, most commercial MT systems produce better translation than so called on-line services. To propose a new approach that results in achieving almost human-like quality translation, we decided to analyze one of the most popular tool Alta Vista Babel Fish service based on the Systran system

1. Related Work

Since 1933, when Georges Artsrouni from France and Petr Trojanskij from Russia made the first ever attempt to construct mechanical multilingual dictionary with implementation of Esperanto-based symbols used for coding and interpreting grammar rules [1], the rapid technological advances in the field have been revolving around the question of building software that produces human-like translation quality.

In 1954, at Georgetown University there was a presentation of a GAT fully automated system that translated sixty sentences from Russian into English with the satisfaction rate of 92% reported in 1972 [4]. The translation process was based on six grammar rules and 250 vocabulary entries and worked on an IBM computer.

The first approaches to computer-assisted translation developed are:

- direct model – word-for-word,
- transfer - based model – added syntax rules that eliminated ambiguity,
- interlingua model – a universal knowledge representation of tokens.

Over the last decades, depending on the language pair co-relations, which is specific in each case, the models have been modified.

Methodology for organizing translation process relies on language resources (dictionary-based), example-based (parallel corpora) and statistical approaches, which are found the most effective. In our approach we combine them all [5], [7]. The techniques developed at the beginning are still expanded by other research groups. The examples include exploiting example-based MT [15], interactive translation [16], using part-of-speech tagging [17], or N-grames [14] or one of the earliest techniques called controlled vocabulary [13].

Despite the dramatic progress in the field, still most MT systems produce gist translation called automatic, or fully automatic high quality translation (FAHQT) [6]. Babel Fish is an example of the first text translation publicly available on the Web in 1997. European Community set up a research project to build English-French version of the Systran system [2]. The first version of Systran was programmed in macro-assembler language soon after completing the GAT project [4].

Nowadays, the trend is towards building multilingual MT systems with advanced search engines that browse the Web for the documents in a language other than the query language like e.g. Alta Vista.

2. The test Set Structure

For the purpose of this study we constructed a test set that consists of 31 grammar structures each of which has been divided into two parts: one or more example sentences and possibly lots of sentences that include the same structure extracted from the French version of the Europarl corpora.

The structure of our test set covers four main grammar areas: lexical devices, morphology, coherence and cohesive devices [10] extracted from the corpora. The critical point at this stage of our experiment was to adopt such a methodology in each case that will result in making the subgroups absolutely separate from each other. Thus, we established at first the approaches to avoid disambiguation.

LEXIS	MORPHOLOGY	COHERENCE	COHESION
ACRONYMS	MORPHEMES	DEFINING CLAUSE	CATAPHORIC REFERENCING
MORPHEMES	AFFIXATION	NON-DEFINING CLAUSE	ANAPHORING REFERENCING
HYPONYMY	COMPOUND WORDS	COMPOUND SENTENCES	LEXICAL COHESION
IDIOMS	CONTRACTIONS	COMPLEX SENTENCES	SUBSTITUTION
FIXED PHRASES		DEPENDENT NON-FINITE CLAUSE	CONJUNCTION
COLLOCATIONS		ADVERB-ADJECTIVE RULES	
MULTI-WORD VERBS			

Figure 1. Grammatical structures of the test set

In order to clarify our approach we define some of the categories [20]:

- acronym – initial letters of words or word parts in a phrase or a name
- morpheme – the smallest linguistic unit that has semantic meaning
- hyponymy – inclusion of a word subgroup
- fixed phrase – a group of words that function as a single unit in the syntax or a sentence
- compound – two words joint with hyphen
- sample sentence – made of one clause
- idiom – an expression or a phrase whose meaning cannot be deduced from the literal definition
- cataphoric referencing – coreference of an expression with another one that follows it
- anaphoric referencing – use of articles to point backwards
- substitution – replacing a noun with a word “one/ones”

3. Translation Results with BabelFish Services

Our test set was submitted for translation to the BabelFish services. Here are some example sentences of the translation process:

(1) IDIOM He worked himself to the bone. - il s'est travaillé à l'os. - It was worked with the bone. (0.164)

(2) FIXED-PHRASE I never understood the ins and outs of his job.- Je n'ai jamais compris les coins et recoins de son travail. - I never understood the corners and recesses if its work. (0.337)

(3) MULTI-WORD VERB They didn't have anywhere to put us up so we stayed in a hotel. - Ils n'ont pas eu n'importe où pour nous mettre vers le haut ainsi nous sommes restés dans un hôtel. - They did not have to put to us anywhere to the top thus we remained in a hotel. (0.427)

(4) COMPOUND kind-hearted – bienfaisant beneficial (0.418)

Here the first sentences are the examples of the categories labeled. They are the reference English into French translations whereas the last sentences are their hypothetical French into English translations.

With regard to the Europarl corpora, the reference sentences have been produced by professional French into English human translator and then selected accordingly while the hypothetical sentences have been translated by the BabelFish on-line services. The numbers in brackets are the scores given by the metric for the whole category that contains additionally the sentences extracted from the Europarl corpora.
Here are the examples from the Europarl corpora:

(1) As you have noticed the great “bug of 2000” has not come true. - Comme vous avez pu le constater, le grand " bogue de l'an 2000 " ne s'est pas produit - Although, as you will have seen, the dreaded - 'millennium bug' - failed to materialize.

(2) Still, on the other hand, the citizens of' a certain number of our countries were victims of natural disasters which were really terrible. - En revanche, les citoyens d'un certain nombre de nos pays ont été victimes de catastrophes naturelles qui ont vraiment été terribles . - Still the people in a number of countries suffered a series of natural disasters that truly were dreadful.

The example sentences presented above show how much the translation process is irreversible. This is a starting point for making improvements both in the translation algorithm and the language resources inbuilt.

In the next step, each category was evaluated with Meteor, the metric that matches words, called unigrams, of hypothetical translation and its reference translation [9]. This metric is known as more precise than Blue [8] and closer to the human judgment. It has been used during the TREC and NIST campaigns over many years as the most precise metric for evaluation of translation quality.

Provided that there are more reference translations, only the best score is reported. The word alignment relies on comparison of strings between these two translations:

- The Exact module, in which only the same words are compared
- The Porter stem module that compares only the words’ stems
- The WN synonymy module compares synonyms of the word pair

As a result, Meteor produces Precision (P=mu/tu) as a proportion between the number of mapped unigrams *(*mu) to the total number of unigrams *(*tu) in the translation and Recall (R=mu/turef) – a ratio of mu to the total number of unigrams in reference translation (turef). The system creates chunks that are the units of two matched unigrams of the same order and calculates a harmonic mean (F-mean*),* fragmentation penalty (Pen) using a fragmentation fraction denoted by Ch/mu, and a score (Sc) for alignment as follows:

$$F - men = \frac{PR}{\alpha P + (1-\alpha)R} \quad (1)$$

$$Pen = \delta * (\frac{Ch}{mu})^{\beta}$$

$$Sc = (1 - Pen)\frac{Ch}{mu}$$

Ch – number of chunks

α, δ, β – parametres as described in [8]

Parameters α, δ, β, being the weights of F-mean, fragmentation function and fragmentation penalty in the equation above, were initially assumed as 0.9, 3.0 and 0.5 respectively, but after some more experiments (version 0.6), they have been optimized to make the overall score closer to the human judgment. The intention of creating the penalty function was to deal with longer matches.

3.1. Systran Translation Services

Systran, a producer of BabelFish translation services, uses the Interlingua model in their distributions. The engine translates texts of no more than 150 words long in 19 languages including French.

In 2001, the producer introduced declarative programming that allows the designer to describe some language phenomena by graphical formalism rather than coding the task steps [19]. Interlingua model has been developed to implicit transfer based on parallel source and target descriptions. In addition, XML exchange format component was added. Thus, natural language is supported by the following modules [18]:

- document filter for text extraction and code formalization,
- encoding and character set converter,
- language recognizer,
- pre-processor for identifying document types,
- spell checker,
- sentence segmentation component,
- word delimiter for creating word forms,
- part-of-speech tagger,
- text synthesizer for production word forms,
- semantic domain recognizer for subject ID.

In 2004, a commercial system was tested on Spanish into English translation and scored 0.56 with Blue [6]. Before submitting our test set, we compared some of the translation results produced by Systran Professional 3.3 [11] to BabelFish services. Surprisingly, more than 90% of the French sentences was translated into English in the exactly the same way which indicates that the technique of processing the text has not been changed. Some problems like different use of separators by the engine, document structure that is not preserved by the MT systems, or poor parser performance cause the same errors in both systems compared [12]. Obviously, the commercial Systran

distributions are provided with a range of facilities not included in the on-line translation services.

3.2. Lexis measured with Meteor

Our test set consisted of 31 examples (texts, word lists, or sentences), each relating to a different feature, scored 0.7105 with the Porter stem module and 0.6935 with the Exact module, whereas the sentences alone extracted from the Europarl corpus, scored 0.5330 with the Porter stem module and 0.5231 respectively.

However, using the newest version of Meteor, our test set scored 0.6663 and the Europarl sentences 0.5168 with the exact module only. The difference is a result of replacing constant parameters with the variable ones in order to make the evaluation closer to the human judgment.

Due to the fact that Meteor reports only the best score out of all the segments assessed, we divided the test set into subsets of sentences belonging only to one category of the discourse analyzed. The aim was to find the features that are translated with the highest accuracy and those with the lowest one.

Table 1. Evaluation of the translation quality of the lexical features

LEXIS	Acronyms	Acronyms	Morphemes	Hyponymy
Score	0.121	0.315	0.528	0.463
Matches:	2.000	3.00	4.000	4.000
Chunks	2.000	2.000	2.000	2.000
HypLength:	10.00	9.000	8.000	9.000
RefLength	8.000	8.000	7..000	8.000
Precision:	0.200	0.333	0.500	0.444
Recall:	0.250	0.375	0.571	0.500
1-Factor:	0.222	0.352	0.533	0.470
Fmean:	0.243	0.370	0.563	0.493
Penalty	5.000	0.148	0.062	0.062
Fragmentation	1.000	0.666	0.500	0.500

LEXIS	Idioms	Fixed phrases	Collocation	Multiword verbs
Score	0.164	0.337	0.490	0.427
Matches:	3.000	4.000	5.000	10.00
Chunks	3.000	2.000	3.000	8.000
HypLength:	10.00	12.00	10.00	21.00
RefLength	9.000	11.00	9.000	17.00
Precision:	0.300	0.333	0.500	0.476
Recall:	0.333	0.363	0.555	0.588
1-Factor:	0.315	0.347	0.526	0.526
Fmean:	0.329	0.360	0.549	0.574
Penalty	0.500	0.062	0.108	0.256
Fragmentation	1.000	0.500	0.600	0.800

The table shows a difference between evaluation of the text as a whole and its crucial points in particular. We observe that the features which occur the most often achieve relatively better results than the others. Therefore, the highest score goes to the morphemes and the lowest one to the acronyms and idioms.

3.3. Morphology measured with Meteor

This section presents distribution of morphological features. To be consisted with the ESOL project, we decided to include morphemes again, but this time in relation to morphological features. Since they are in fact a part of affixation so that to make them a separate feature, we agreed to exclude the suffixes from it.

Table 2. Evaluation of translation quality of the morphological features

MORPHOLOGY	Morphemes	Affixation	Compound words	Contractions
Score	0.5282	0.4259	0.4189	0.1639
Matches:	4.0000	6.0000	3.0000	1.0000
Chunks	2.0000	4.0000	2.0000	1.0000
HypLength:	8.0000	12.000	7.0000	7.0000
RefLength	7..000	12.000	6.0000	6.0000
Precision:	0.5000	0.5000	0.4286	0.1429
Recall:	0.5714	0.5000	05000	0.1667
1-Factor:	0.5333	0.5000	0.4315	0.1538
Fmean:	0.5634	0.5000	0.4315	0.1639
Penalty	0.0625	0.1481	0.1481	0.0000
Fragmentation	0.5000	0.6667	0.6667	1.0000

Despite a lack of contractions which have been removed from the text, a few extracted from the whole corpora was translated by the BabelFish services reasonably well like e.g. couldn't, or don't.

3.4. Coherence measured with Meteor

On the contrary to cohesion, coherence addresses semantic meaning of the sentences.

Table 3. Evaluation of translation quality of the coherence features

COHERENCE	Defining Clause	Compound Sentences	Complex Sentences	Main Clause	Simple Sentences	Dependent Nonfinite Clause	Nondependent Clause
Score	0.415	0.810	0.614	0.747	0.225	0.371	0.606
Match	10.00	18.00	8.000	26.00	5.000	5.000	9.000
Chunks	8.000	4.000	4.000	8.000	5.000	3.000	4.000
HypLen	17.00	23.00	14.00	37.00	12.00	12.00	16.00
RefLen	18.00	22.00	12.00	34.00	11.00	12.00	14.00
Precision	0.588	0.782	0.571	0.702	0.416	0.416	0.562
Recall	0.555	0.818	0.666	0.764	0.454	0.416	0.642
1Factor	0.571	0.800	0.615	0.732	0.434	0.416	0.600
Fmean	0.558	0.814	0.655	0.758	0.450	0.416	0.633
Penalty	0.256	0.005	0.062	0.014	0.500	0.108	0.043
Fragm	0.800	0.222	0.500	0.307	1.000	0.600	0.444

Both complex sentences as well as dependent non-finite clauses achieved the highest score since they are characteristic to formal utterances presented in the European Parliament. However, almost all the features were processed correctly by the system. Furthermore, we grouped the sentences according to some adverb-adjective rules. The next figure shows the following of such rules:

1. Qualitative + classifying adjectives.
2. Predicative use of the adjective.
3. Irregular adverbs.
4. Participle.
5. Qualifiers.
6. Extreme adjectives as intensifiers.
7. Adverb-adjective cross-references.

Table 4. Adverb and adjective rules of coherence evaluation

COHERENE ADV-ADJ RULES	Adj-Adv 1	Adj-Adv 2	Adj-Adv 3	Adj-Adv 4	Adj-Adv 5	Adj-Adv 6	Adj-Adv 7
Score	0.335	0.537	0.760	0.597	0.391	0.514	0.636
Matches:	5.000	5.000	10.00	5.000	5.000	10.00	17.00
Chunks	4.000	2.000	2.000	2.000	3.000	4.000	8.000
HypLength	12.00	9.000	14.00	9.000	15.00	17.00	28.00
RefLength	11.00	9.000	13.00	8.000	11.00	19.00	25.00
Precision	0.415	0.555	0.713	0.555	0.333	0.588	0.607
Recall	0.454	0.555	0.769	0.625	0.454	0.526	0.680
1-Factor	0.430	0.555	0.740	0.588	0.384	0.555	0.641
Fmean	0.450	0.555	0.763	0.617	0.438	0.531	0.671
Penalty	0.256	0.032	0.004	0.032	0.108	0.032	0.052
Fragmentation	0.800	0.400	0.200	0.400	0.600	0.20	0.400

For adverb – adjective cross-references, both the reference and hypothetical texts proved the longest. These features were used quite often by the Speakers. Qualitative + classifying adjective rule is characteristic for expressing emotions rather than giving a talk, thus the score is relatively low.

3.5. Cohesion measured with Meteor

Cohesive devises correlate grammatical aspects with lexis of the text. This study does not deal with ellipses as the rarest cohesive devise used by the speakers and too difficult to be recognized by the text analyzers. In natural language, the speaker avoids the words or phrases mentioned before to make the utterance more formal.

Table 5. Evaluation of translation quality of the cohesive devices

COHESION	Cataphoric Referencing	Anaphoric Referencing	Lexical Cohesion	Substitution	Conjunction
Score	0.7397	0.5324	0.543	0.3156	0.685
Matches	9.0000	11.000	5.000	4.0000	16.00
Chunks	2.0000	7.0000	3.000	3.0000	5.000
HypLength	13.000	18.000	10.00	10.000	23.00
RefLength	12.000	18.000	8.000	10.000	23.00
Precision	0.6923	0.6111	0.500	0.4000	0.695
Recall	0.7500	0.6111	0.625	0.4000	0.695
1-Factor	0.7200	0.6111	0.555	0.4000	0.695
Fmean	0.7438	0.6111	0.609	0.4000	0.695
Penalty	0.0055	0.1281	0.108	0.2109	0.015
Fragmentation	0.222	0.636	0.600	0.750	0.312

Cataphoric referencing is not that common as the anaphoric one as it is a way of introducing a subject in an abstract way and then addressing it directly. For substitution

we extracted words like “one/ones”. Lexical cohesion was based on definite articles and determiners. Grammatical conjunction scored high as it is correctly translated by the most MT systems.

4. Reliability of the Discursive Features

In this section we analyze reliability of features in correlation to their co-occurrences in the Speakers’ turns.

Reliability is defined here as an absolute value of a distance function of the co-occurrences in the corpora to the translation accuracy.

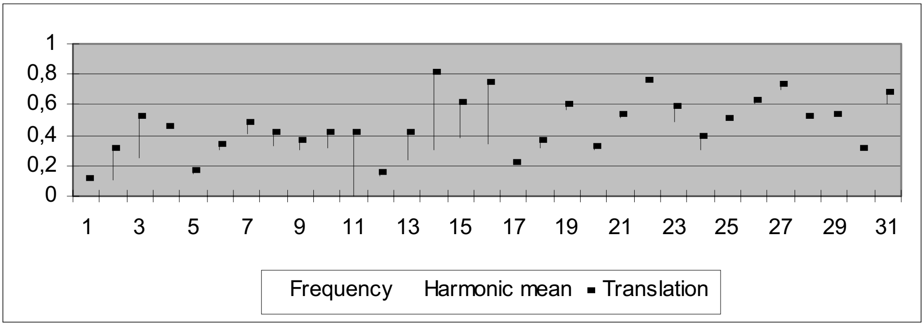


Figure 2. Distance-based model of co-occurrences of the discourse features and the translation accuracy

The diagram presents the distance that is defined as an absolute value of frequency of the features and their translation quality measured with Meteor. The horizontal axis shows features’ numbers 1 to 31 and the vertical axis, their values taken from the tables above. The length of the vertical lines indicates the distance.

We observe that for most features the occurrence is proportional to the translation quality. The only exception is a coherent device, in particular adjective-adverb rules. The harmonic mean is a weighted average of precision and recall calculated by Meteor known as 1-Factor.

5. Information Retrieval of the Discourse Features

In the next step, all the translation results were submitted to AltaVista search engine for information retrieval. At the beginning, we decided to concentrate only on the Web and News search types leaving out Images, MP3/Audio and Video searches, but because the number of News is close to 0 in most cases, we abandoned the search type as well.

Each category of English phrases was submitted separately to retrieve information in French. Relevant items were only those that included the whole query (in quotes as an exact query). AltaVista records the number of results below its search box.

For each query we calculate precision (*Pr*) and recall (*R*) since the goal is to evaluate the correlation between translation quality and the number of the relevant search results retrieved on the Web. The same procedure is followed for both the reference and the hypothetical test sets.

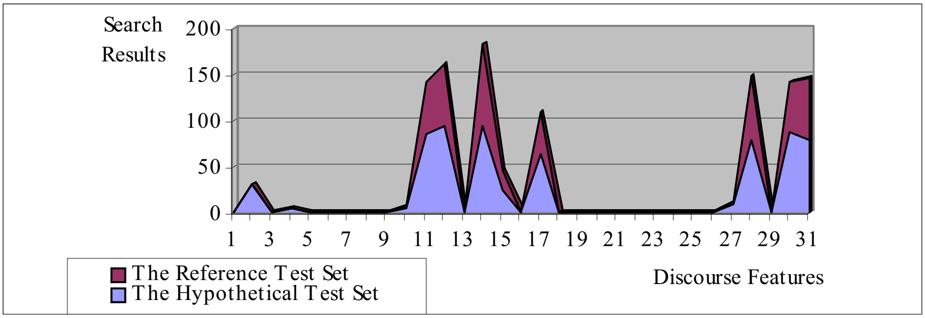


Figure 3. Comparison of the search results for the Reference and Hypothetical Test Sets

The graph shows for which features the number of AltaVista responses drops significantly as a result of the incorrect translation quality. Furthermore, we noticed that for the features with relatively low number of the system responses, like features 3-9, 18-26 and the seven adjective-adverb rules, the limit of distance between translation accuracy and their occurrence goes to 0. Another finding indicates the exception for morphemes – in both cases, that is the reference and the hypothetical test sets, the number of search results from AltaVista was about the same. The analysis carried out in section 2 confirms this finding.

6. Conclusions

In our test sets we collected examples of the features from the ESOL project and lots of the sentences from the Europarl corpora. Although backward translation is not a novel approach, here it is used to show how to test particular system components in order to identify the reason of incorrect translation results.

Our findings show the methodology that deploys irreversibility of the translation process aimed at evaluation of the system performance. Also, despite the formal register required in the European Parliament, the features that occur the most often in the Speakers' turns, are very popular on the net as well.

Another finding indicates that not all of the linguistic structures analyzed here impede the translation process e.g. coherence devices, adjective-adverb rules in particular. However, it is an essential issue to relate the factors that influence the limitation of the system responses but only those that cover the area of the user's interest.

The last, and presumable the most surprising result of the research presented, shows that in case of Systran known as the most popular producer of machine translation products, the commercial distribution gives the same results in 93% that the on-line services AltaVista BabelFish.

Furthermore, the impact of translation quality on retrieval results indicates what aspects of language need to be studied to make the Internet cross-lingual.

As seen, it seems a " big hole" for the research community to continue project aimed at determining the situations and the impact factors like e.g. translation accuracy that disrupt or not the overall performance of the cross-language information retrieval systems.

Our method is efficient and can be ported to other language pairs. In our further study we plan to concentrate on the translation models in order to analyze the performance of the system components to the linguistic features established here in particular.

References

[1] J. Hutchins, Machine Translation History, *Encyclopedia of Language and Linguistics*, Second Edition, Elsevier, vol. 7 pp.375-383, Oxford, Elsevier, 2006.
[2] M. Wapier, The Soldiers are in the Coffee – an Introduction to MT, *Cultivate Interactive*, issue 2, 2000.
[3] J.White, Toward an Automated Task-Based MT Evaluation Strategy, *Proceedings of the Workshop on MT Evaluation at LREC*, 2000.
[4] Ch. Boitet, Factors for Success (and failure) in MT, *Fifth MT Summit*, Luxemburg, 1995.
[5] Ch. Munning, H. Schutze, Foundations and Statistical NLP, *The MIT Press*, Cambridge, Massachusetts, London, 2000.
[6] E.Ratliff, Me Translate Pretty One day, *WIRED*, issue 14.12, 2004.
[7] D. Geer, Statistical MT Gains Respect*, IEEE Computer*, 2005.
[8] C. Callison-Burch, Re-evaluating the role of BLUE in MT Research, *11th Conference of the European Chapter of the Association for Computational Linguistics*, CACL 2006.
[9] S.Baneriee, A. Lavie, Meteor: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and Summarization at the 43th Annual Meeting of the Association of Computational Linguistics*, Michigan, 2005.
[10] M. Halliday, H. Rugayia, *Cohesion in English*, Longman, 1976.
[11] D. Lewis, PC-Based MT; An Illustration of Capabilities in Response to Submitted Test Sentence, MT Review No 12, *The Periodical of the Natural language Translation*, Specialist Group of the British Computer Sociaty, Issue No 12, 2001.
[12] P. Senellart, *Systran Translation Stylesheet*, 2000.
[13] Takako Aikawa, Lee Schwartz, Ronit King, Mo Corston-Oliver, & Carmen Lozano: Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. *MT Summit XI,* 10-14 September 2007, Copenhagen, Denmark. Proceedings; 1-7.
[14] J.M. Crego, A. de Gispert, P. Lambert, M.R. Costa-jussà, M. Khalilov, R. E.Banchs, J.B.Mariño, & J.A.R.Fonollosa, N-gram-based SMT system enhanced with reordering patterns. HLT-NAACL 2006: *Proceedings of the Workshop on Statistical Machine Translation*, New York, NY, USA, June 2006; 162-165.
[15] M. Hearne, A. Way, Disambiguation strategies for data-oriented translation. *EAMT-2006: 11th Annual Conference of the European Association for Machine Translation*, June 19-20, 2006, Oslo, Norway. Proceedings; 59-68.
[16] J. Tomás, F. Casacuberta, Statistical phrase-based models for interactive computer-assisted translation. Coling-ACL 2006: *Proceedings of the Coling/ACL 2006* Main Conference Poster Sessions, Sydney, July 2006; 835-841.
[17] C. Lioma, I. Ounis, Deploying part-of-speech patterns to enhance statistical phrase-based machine translation resources. *ACL-2005: Workshop on Building and Using Parallel Texts – Data-driven machine translation and beyond*, University of Michigan, Ann Arbor, 29-30 June 2005; 163-166.
[18] M. Flanagan, S. McClure; Systran and the Reinvention of MT, *IDC Bulletin* #26459, January 2002.
[19] J. Senellart. C. Boitet, L. Romary; Systran New Generation, *MT Summit* IX; September 22-26, 2003.
[20] M.Swan: *Practical English Usage*, Oxford University Press, 2000.

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-182

Modelling Agent Behaviours in Simulating Transport Corridors Using Prometheus and Jason

Karol KAIM and Mateusz LENAR
Wrocław University of Technology, Institute of Applied Informatics
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
e-mail: mateusz.lenar@pwr.wroc.pl

Abstract. This article describes the approach to create the multi-agent system (MAS) for simulating transport corridors. There are few new ideas that were used during the developing the system. JASON language used in implementation of this system is a relatively new approach. A commitment based communication was described before in few papers but there is a little about its implementation. The social aspects of agents are emphasizes in this paper. Although the system is not fully implemented, the combination of Prometheus methodology and JASON language seems to be the good choice for developing the multi-agent systems.

Keywords. agent technology, modelling agent behaviours, transport corridor simulation, multi-agent systems

Introduction

Agent technology is recognized as a suitable approach to develop complex, open and distributed systems. Although multi-agent systems are constantly studied, they are still rather experimental method. Recent studies focus highly on social aspect of the system.

Section 1 provides the description of transport corridor which will be simulated by multi-agent system. Section 2 presents the problem and proposes one of the possible solutions. Section 3 presents the overview on multi-agent systems and multi-agent based simulation. Section 4 proposes including approaches known from social science into agents. Section 5 summarizes known approaches to the design of agent communication and proposes the way of implementing the commitment based approach into the system. Section 6 compares briefly most common agent methodologies and in Section 7 Prometheus methodology is presented. Section 8 describes JASON language, which was recently developed. Section 9 presents the overview of the system. The 10th Section contains conclusions and describes future work.

1. Transport Corridors

A transport corridor is a route that connects two points (usually cities which, de facto, implies connecting of different regions) – it can be a roadway, railway, sea, or air

connection. A good example of a transport corridor is the historical Silk Road, which connected China and Europe. Thus, transport corridors have helped establishing trade routes and have an impact on a whole society, such as in case of globalization. An interesting suggestion of PJ O'Rourke is presented in [1]: “it wasn't so much the presence of a whole pile of nuclear ordinance that had caused the bricks [of Berlin Wall – author] to fall as the availability of the Sony Walkman.”

Establishing transport corridors not only improves the international exchange of goods and opens new markets, but also transport corridors can have a great impact on a local region – [2] even suggests that influence on a domestic transport may be greater than on international one.

The usual transport corridors are highways or rail tracks between important trade centres. Often new routes are built near the existing ones in order to limit pollution. In recent time, keeping pollutions and gas emissions as low as possible has been seen important not only by ecological groups but also by scientists, e.g. [3]. However, the most interesting area seems to be the intermodal transport corridor where sending goods between the origin and the destination can be divided into sending goods between different “nodal points” (in most cases terminal – points where cargo is reload on the other type of transport). Furthermore, as [3] points out intermodal transport may be much more ecological.

In research of Henesey and Persso, [4] they propose an interesting model for simulating decision making on a transport corridor. In the model the consider using six agents in simulating a generic transport corridor: freight forwarder, governmental-legal authorities, inland transportation provider, shipper, shipping lines and terminal. In this model they do not consider the transportation cargo via air – using airplanes, which is both extremely fast and expensive (thus it is only suitable for a very specific type of cargo). Description of actors is as follows (based on [4]):

Freight forwarder – freight forwarder is responsible for taking all necessary actions (e.g. checking with the government-legal authorities) to transport goods from origin to the destination for a shipper. Freight forwarder is usually used when the shipper lacks of own shipping department.

Governmental-legal authorities – customs and governmental agencies. They influence the transportation of good in two ways. First of all, they make policies including taxes and subsidies. Secondly, they are responsible for inspections and insuring that sending goods is according to the law. Policies and the way in which there are processed influence choosing particular transport corridor (e.g. choosing Finnish over Russian ports [5])

Inland Transportation Provider – inland transportation provider is responsible for transportation cargo via land – goods may be shipped by train or by truck. What is more, land transportation may include different types of vehicle at some points of transportation (e.g. in some area or at some time truck traffic may be forbidden and cargo or even trucks are transported by train). However, in most cases inland transportation provider would have only one type of fleet (i.e. trucks or trains).

Shipper – shipper is a person or organization that is willing to send goods via transport corridor. The shipper may be an importer, an exporter or both. It is supposed that bigger organizations can negotiate better transport conditions – it may leads to establishing an organization that unites small companies, e.g. Swedish log companies [4]. The shipper is supposed to have the most influence on traffic in the transport corridor.

Shipping lines – shipping lines are organizations that transport cargo between seaports. Transporting goods on ships is usually only one phase of the whole process. Need of integrating sea and land transport has already emerged and shipping lines started developing integrated logistic systems.

Terminal – terminal is a point where goods are either unloaded from ships and send further by inland transport providers or loaded on ships. Nowadays, terminals are not the static point of exchanging goods, but act as a 'nodal points'.

Above six actors creates, in the context of simulation, a society. Integrating their efforts (especially shipping lines and inland transport providers) and coordination their actions is of great importance – not only does it improves the efficiency of transport corridor itself, but also provides benefits for local communities [2] and may decrease the pollution [3].

2. Description of the Problem

2.1. The flow of Goods in Transport Corridors

Transport corridor is the way that connects two points. Although transport corridors used to be considered as a one route, nowadays there are rather designed area on which transportation is possible. Modern transport corridors are often intermodal, i.e. there is possibility to use different kind of transport on different stages. Moreover, there is usually also possibility to choose a slightly different route within a corridor. For an instance, in order to transport goods from Stockholm, Sweden to Wroclaw, Poland one may choose to use ferry from Ystad to Swinoujscie, Poland and then use a train from Swinoujscie to Wroclaw, while it is also possible to transport cargo using trucks from Stockholm to Karlskrona, Sweden, load a truck on ferry to Gdansk, Poland, and then continue transporting it by truck to Wroclaw.

The flow of goods starts when one (shipper) decides to transport the goods. Shipper can either a) transport it on its own, or b) organize the whole transportation, or c) ask a freight forwarder to transport it. The decision depends on many factors. Tourists prefer to transport souvenirs on their own as well as some organization with transport divisions, some organizations uses freight forwarders, while some (e.g. small one that tries to cut down costs) prefer to organize the whole process on its own. If a shipper turns on to the freight forwarder, she\he has no more duties but to check if a good arrives safely. To organize the transport of a good, the shipper or the freight forwarder must choose the route, i.e. inland transportation providers and/or shipping lines. It may also be necessary to contact with governmental-legal authorities (e.g. while moving cargo across borders) and terminals (to reduce the number of transactions while handling cargo at terminals and stations connected with it).

2.2. Measures of Transporting Goods

More options of transportation goods lead to emerging of need to estimate advantages and disadvantages of choosing a specific solution. To be able to estimate the profit of any particular choice, some metrics should be presented. Two measures which are tightly connected with transporting are time and cost. Time can be considered as a continuous or discrete (i.e. if a good is transported within a time frame). In spite of the fact that cost is usually associated with economic cost, it is broadly speaking rather

transaction cost. In [6] nine types of cost related to the choice of transport corridor are proposed. Another measure may be trustworthiness – it is possible to use different ways of sending goods when it is important to assure that it would arrive on a specific day without any damages to a good itself. Combination of these three measures (with specific weights for specific shippers) seems to be a reasonable metric for choosing the right way of sending goods.

2.3. Simulation

The flow of goods in transport corridor is be simulated in order to choose the best alternative of transport and observe the influence of different factors (e.g. changes in policies). In this case, the simulation is preferred over optimization due to the complexity (i.e. many different organizations negotiate the terms of transport). Simulation also enables observing the differences in results caused by choosing different policies or acting in different environments.

2.4. System

The proposed system should enable simulation of complex environment. Agent technology seems to be suitable for this problem. Multi-agent systems are considered to be a good solution for complex (many different organizations and need of negotiations between them) and open (there is possibility of adding a new company/shipper) systems. What is more, the multi-agent system is a good metaphor of a whole environment (each actor may be an agent). Multi-agent systems are also believed (and, in fact, it was proved) to be suitable for simulating complex environments. There are many advantages of using multi-agent based simulation. First, it enables more intuitive and appropriate way of modelling. Real organization may be represented by an agent. Actors, which are independent and autonomous, may be easily modelled as agents – they not only react to changes of the environment, but also may perform proactive actions. Moreover, in opposition to macro simulation, the observation of a particular agent and its locality may be performed. Furthermore, each agent may evolve and adapt to the environment. Finally, multi-agent system is very flexible – new organization (agents) may be easily added and change of behaviour and interaction of agents can be observed.

The system will be designed using Prometheus methodology and implemented using Jason.

3. Multi-Agent Systems

Multi-agent systems are becoming more and more popular. One of the reasons of this is increasing complexity of systems and ease with which a robust system can be designed. [10] even states that "Multiagent systems are the best way to characterize or design distributed computing systems". However, [8] suggests that only some class of problems should be solved with multi-agent systems (what is more it notes that agent technology is in some cases abused).

One of the most common and simple definition of a multi-agent system is an environment that contains multiple agents. However, the agents create not a set, but a society. One of the most accurate definitions of an agent is presented in [8]: "an agent

is a computer system that is situated in some environment and that is capable of autonomous action in this environment in order to meet its design objectives." This definition may not capture all of the agent properties, but as Russel and Norvig (see [9]) said: "The notion of an agent is meant to be a tool for analyzing systems, not an absolute characterization that divides the world into agents and non-agents."

To perform the simulation of transport corridor the multi-agent system has been chosen. However, there might arise two questions: firstly, why bother with simulation? Secondly, does multi-agent systems posse any advantages over other methods? The reasons for simulation are numerous (see [10]):

Cost – it is usually considered to be less expensive to build a model and perform the simulation instead of experimenting with the real system

Time – even thought that building a simulation model may require a lot of time to build, a single run of simulation is by far shorter than obtaining the results from the real system. Simulation may be repeated multiple times in order to gain more accurate information. On the contrary, it is also possible to run simulation slower than actual system when it may be beneficial (i.e. when real systems act to fast to be able to observe it)

Repeatability – possibility to run the simulation in exactly the same way several times might be desired in order to investigate the factors for the particular result. Moreover, changing only factor might give some insight on its importance. This is not possible in real system as the conditions are changing constantly

Safety – modelling the dangerous situation (e.g. earthquakes) allows simulating it without any danger. Moreover, performing simulation in safe laboratory environments allows making mistakes that might be otherwise hazardous

Modelling a transport corridor and simulating the flow of goods brings some benefits. It does provide tool, which can be used to observe what the effects of a particular change are. The simulation may also present the solution which route should be taken and which companies are supposed to be the best partners.

Though the most common way of simulation used to be the event-driven simulation, the complexity of modern system usually are troublesome to present this way.

"The use of agent systems to simulate real world domains may provide answers to complex physical or social problems that would otherwise be unobtainable due to the complexity involved, as in (...) modelling the impact of public policy options on social or economic behaviour" [11].

The multi-agent based simulation (MABS) was successfully implemented previously. However, the multi-agent systems are still rather young. The main drawback of this technology is that methodologies and languages are still experimental. Furthermore, lack of tools (at least the completed designing environments) and lack of debuggers makes it more difficult to implement a system and raise the possibility of bugs. Moreover, the lack of supporting methods and tools increase the costs. Nevertheless, the multi-agent systems (which can be seen as a metaphor of a human organization – compare [8]) seems to be the most suitable solution for this problem.

4. Social Science and Agents

The prisoners' dilemma and the Axelrod's tournament (see [8]) prove that the best choice available from rational point of view is not always the best in a long

perspective. The sad conclusion that might be drawn is that the best choice is irrational. However, it might not be the case. In fact, people sometimes behave "irrational" – e.g. helping their friends. Although it does not is connected strictly with any benefits (except any potential benefits in future which cannot be linked straight to this action), it is usually consider as "right" in a society. In some situations cooperation (meaning the work for the common good) may be fruitful in the future. The question that arises is how this kind of thinking may be implemented into agents? As TIT-FOR-TAT and Axelrod's tournament shows, building the complex model of others might not be the solution. Not only it was not successful when the problem was well described (as in prisoner's dilemma), but it also might be impossible to build the model of other in an open environment.

Before deciding how an agent should react in a society, it might the good idea to look on what social science can say. First thing which should be mentioned is that people who live in a society are aware of themselves and other – furthermore, there create own "self". As Charles H. Cooley wrote: "Self and society are twin-born." Thus, it is not possible to create a social agent without consideration of a society. Cooley also states that self creates in interactions with others (I am what I think you think of me). Even more interesting is a work of George H. Mead. According to [7], he divided self into reflexive self – called by Mead "me" - and spontaneous self – called by Mead "I". "I" is a unique individuality, which creates reactions on the others' actions. "Me", on the other hand, is a picture of reactions, judgements and expectations. Mead also distinguish others from "important others" – people who are more important for a specific person.

As far as agent technology is concerned, the agents always have the "I" part of self – they have some implemented mechanisms of behaviour (e.g. believes, desires and intensions as in BDI architecture). However, there is a lack of "me" part – agents conduct actions because, as M. Wooldridge in [8] states, "they want to". Those actions are oriented on increasing own welfare. Besides the own welfare the wellbeing is also important for people. Actions oriented on making a good impression on "important others" (e.g. helping a friend) may seem irrational for an agent (no explicit payback for taken effort) but are understandable for (at least most) human beings.

5. Communication in Multi-Agent Environment

Computer systems contain multiple entities that require communication among themselves. In object-oriented systems objects exchange messages (that, in fact, can be very complex): it can be either sending some data or even request for a specific action. It is common that objects execute the methods of another object. However, objects do not have any control over the execution of public methods. Nevertheless, this approach was very successful in a wide range of application. Unfortunately, this simple way of communication is no use for agent environments. Agents are considered autonomous and thus can not "be forced" to execute action. The way of communication requires to be extended.

Communication between agents, as [8] indicates, has been recognized as a topic of enormous importance from the very beginning of agent technology. Studies in this field led to development of agent communication languages (ACLs). There are some essential requirements that ACL must meet (see [8], [15]):

- ACL must be verifiable;

- meaning of communication must be expressed in objective and external way;
- ACL must be extensible;
- ACL must be simple;
- ACL must be expressive.

Verifiable of semantic is regarded as deciding whether the system conforms to ACL ([8]). The second property means that meaning of communication is independent on internal agent state. Extensibility allows agents and designers manage with new situations. Simplicity is required to let the designers understand and correctly interpret ACL. The last property assures that ACL can be successfully used in all common scenarios.

As [16] states, there are three major approaches to ACL semantic:

- mentalistic semantic,
- protocol-based semantic,
- social semantic.

That entire semantic basis on speech act theory. Mentalistic semantic stressed the notion of speech acts. It was used in popular Knowledge Query and Manipulation Language (KQML) and FIPA ACL (standard of Foundation for Intelligent Physical Agents). However, as it will be presented later, mentalistic semantic has some major drawbacks. Protocol-based semantic define the meaning of speech act with help of protocol, so the meaning can differ between protocols – this makes this approach unsuitable for open environments. As protocol-based semantic highly depends on a protocol (which can be different in each agent system) and is not general, it will not be discussed in details. Social semantic, which becomes the most popular (see [16]), is based on social commitments and thus the interaction between the agents.

Social semantic extends mentalistic semantic with commitments. According to [17], commitment can be defined as a norm:

"(...) to assert a proposition counts as taking on a commitment to subsequently defend the proposition if challenged" [17].

It is assumed that all rational agents recognize this norm. It is considered that all rational agents possess goals, pursuit them actively, hence have reasons for their actions and can express those reasons. One important idea connected tightly with commitments is commitment store introduced by Hamblin. This store keeps tracks of all commitments done by all participants in dialogue.

Commitments are obligations about truth of some value or promise to take some actions. Commitments are usually considered to have properties like conditions (which must be satisfied to fulfil the commitment), debtor, creditor and content. As [16] suggests, the set of speech act should be as follows: Assert, Argue, Declare, Question, Request, Challenge, Promise, Retract. All but Retract are commonly used in ACL semantics. Assert is used to present information about state of the world. Argue allows agent to present arguments to support claims. Declare is a speech act which is supposed to change the state of world. Question and Request are similar acts: the former asks for the information while the latter requests to change the (specific) value to true. Challenge is used in order to acquire an explanation for some information. Promise enables agent to commit itself to some goal or actions. Retract is considered to be metalevel act that allow agent to withdraw from commitment. All of those speech acts have different violation criteria which enable to decide if commitments are violated. In case of violation, the agent that violated the commitment is punished. However, the kind of punishment is not defined. The paper [16] suggests that it should depend on

type of violated commitment and paper [17] proposes two types of punishment: ostracism (i.e. not accepting messages for time period) and blacklisting (i.e. broadcasting details of offence to other agents). For purpose of the simulation, the following solution is proposed:

- weight of offence depends on type of a violated commitment;
- punishment is the decrease of trustworthy, reputation and reliability of an agent;
- punishment is counted using time window.

This solution is supposed to punish most agents that tend to break their promises while still being capable of forgetting the sins of past.

6. Methodologies

Agent technology, seen as a good solution for complex problems, is a relatively new way of implementing systems. One of the crucial steps for developing multi-agent systems is choosing the right methodology. As [11] states, "there are difficulties in implementation phase, partly due to the lack of maturity in both methodologies and programming tools". Another obstacle for developing agent-oriented systems is:

"In general, integrated development environment (IDE) support for developing agent systems is rather weak, and existing agent tools do not offer the same level of usability as state-of-the-art object-oriented IDEs" [11].

In order to choose the suitable methodology for this work, several different approached on modelling the multi-agent systems was compared. The most common methodologies in literature are MaSE, Gaia, Tropos, and relatively new, Prometheus. All of them have some specific property, but for the purpose of the simulation of transport corridor Prometheus seems to be the most suitable. Availability of tools that support designing and focus on a detailed design is desirable.

7. Prometheus Methodology

The Prometheus methodology was presented in [12] in 2002. It's a detailed AOSE methodology which aims to be easily learned and used even by non-experts (see: [13]). Prometheus divides designing into three phases (see Figure 1): system specification, architectural design and detailed design. There is a sharp focus on supporting detailed and complete develop of intelligent agents. "Regarding the process, only Prometheus and MaSE [among MaSE, Tropos, Gaia and Prometheus – author] provide examples and heuristics to assist developers from requirements gathering to detailed design" [13]. Another advantage of Prometheus is that there are two special tools that support developing agents with this methodology. Being detailed and complete, supported by tools and scalable makes Prometheus a useful methodology. The conclusion drawn from Parandosh comparison [13] presents MaSE and Prometheus methodologies as the most mature and best suited to support agent development. However, Prometheus, in contrary to MaSE, provides support for BDI agent architecture.

Though the authors claims that it "has evolved out of industrial and pedagogical experience", there is no real case (i.e. using methodology not in a controlled or laboratory environment) which would be described in a literature. Nevertheless, it is a small drawback in comparison to advantages of Prometheus.

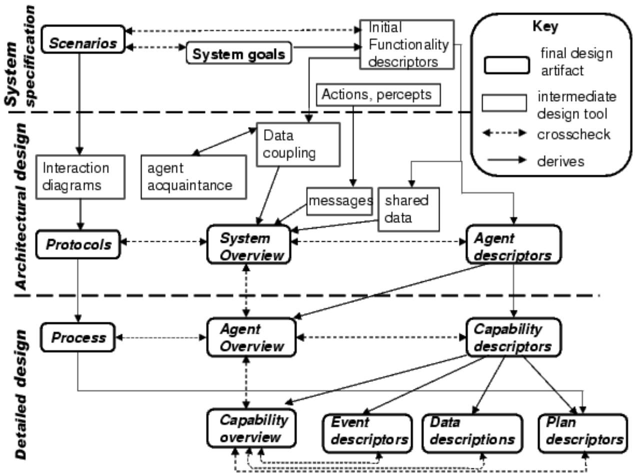


Figure 1. The overview of Prometheus methodology [12]

8. Jason

Agent architecture provides the framework in which agent programs may run. The BDI (Beliefs-Desires-Intensions) agents which have been studied and used for several years are quite popular way of presenting the agents. This architecture gives the agent beliefs, desires and intention. Roughly speaking, beliefs are the information that agent posses (or, at least, in which it believe). Desires are states that agent consider as good for itself. Intensions are the states that agent decided to reach. The practical BDI agents are considered to be reactive planning. They react on the changes in environment, updating their beliefs base and, as a result of some events, taking actions. What the agent should do depends on an agent's intention – actions are supposed to change the environment, so agent need to prepare some plan in order to achieve the desired state.

Jason is the interpreter uses the extended version of AgentSpeak language. AgentSpeak was originally created by Rao (see [14]) and "is a programming language based on a restricted first-order language with events and actions" [14]. Agent is considered to constantly perceive its environment and if any triggering event (i.e. event that trigger execution of some plan) occurs, the suitable plan is added to its intensions (i.e. plans that agent try to achieve). Simultaneously, the agent is reasoning about its current plan and takes actions that aim to change the environment in order to bring the desired state. The key element for Jason is choosing the most suitable plan from plan library. Plan library consist of partial plans – a plan is chosen in result of triggering event and according to the context (i.e. the state of environment and internal state of an agent). Speech-act based communication allows focusing on knowledge-level of communication. Another useful feature of Jason is possibility to customize agent functions and develop simulated environments using Java. Finally, Jason and Prometheus methodology have similar basis and it seems to be a good idea using them both.

9. Model of the System

Model of the system was prepared using Prometheus methodology. Some of the scenarios are presented below (see Figure 2).

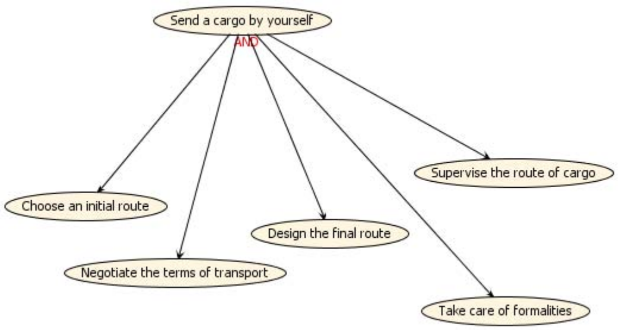


Figure 2 Part of goal overview diagram.

9.1. Scenario: Send a Cargo

Name: Send a Cargo Scenario

Description: As soon as a shipper acquires the new cargo, there is a need to transport it. The shipper must decide on type of preferred transport.

Trigger: The desire to transport a cargo

Steps (no. – type – name – data produced/used):

1 – Percept – Cargo Info received - Produces: Cargo Info

2 – Action – Ask Freight Forwarders for offers – Uses: Cargo Info, Freight Forwarders data

3 – Other – Wait for answers

4 – Percept – Receive offers from Freight Forwarders – Produces: Freight Forwarders Offers

5 – Goal – Choose the most suitable offer – Produces: Offer / Uses: Freight Forwarders Offers

6 – Goal – Negotiate the term of offer – Produces: Final Offer / Uses: Offer

7 – Goal – Assure the cargo was delivered

9.2. Scenario: Propose a Route

Name: Propose a route Scenario

Description: When an inland transport provider or a shipping line receives a inquiry for an offer, it check whatever it has any transport vehicles available for specific time and route and send back the offer.

Trigger: Inquiry from a shipper/freight forwarder

Steps (no. – type – name – data produced/used):

1 – Action - Check availability of transport - Uses: Own Transport data

2 – Goal - Design route - Uses: Own Transport data, Cargo Info / Produces: Route Info

3 – Goal - Estimate the offer - Uses: Route Info / Produces: Offer

4 – Action - Send an offer - Uses: Offer

10. Conclusions and Future Work

Transport corridor as an open and dynamic environment is supposed to be a good candidate for a multi-agent based simulation. Moreover, multi-agent system may be seen as a natural metaphor of this environment. Actors in corridor may be easily presented as agents – they are autonomous and pursuit their own goals.

Combination of Prometheus methodology and Jason language provides the framework for developing intelligent agents using BDI architecture. Emphasizing social properties of agent in communication (i.e. using commitment-based approach) is supposed to increase the value of simulation result.

Further work includes finalizing the implementation of the system and validity simulation results with data obtains from one of the project connected with transport corridors. The final results may be beneficial for both transport corridor and agent community as they may bring some new information about usefulness of current methods.

11. References

[1] "On the road to change", *Commercial Motor* 198 (2003), 66-68.
[2] F. R. Bruinsma, S. A. Rienstra, P. Rietveld, Economic Impacts of the Construction of a Transport Corridor, A Multi-level and Multiapproach Case Study for the Construction of the A1 Highway in the Netherlands, *Regional Studies: The Journal of the Regional Studies Association* 31 (1997), 391-402.
[3] Z. Patterson, G. O. Ewing, M. Haider, The potential for premium-intermodal services to reduce freight CO2 emissions in the Quebec City–Windsor Corridor, T*ransportation Research Part D: Transport and Environment* 13 (2008), 1-9.
[4] L. Henesey, J. A. Persson, Application of Transaction Costs in Analyzing Transport Corridor Organisation structures by Using Multi-Agent Based Simulation, *Promet Traffic & Transportation: Scientific Journal on Traffic and Transportation Research* 18 (2006), 59-65.
[5] *Mivitrans, Intermodal and Transportation Conference*, Germany, Hamburg, 1998.
[6] K. Kylaheiko, D. Cisic, P. Komadina, Application of Transaction Costs to Choice of Transport Corridors, *Economics Working Paper Archive at WUSTL*, 2000.
[7] P. Sztompka, *Socjologia: analiza społeczeństwa*, Znak, Kraków, 2002.
[8] M. Wooldridge, *Introduction to multi agent systems*, John Wiley and Sons Ltd., West Sussex, England, 2002.
[9] S. J. Russell, P. Norvig, *Artificial intelligence: modern approach*, NJ: Prentice Hall, Englewood Cliffs, 1995.
[10] R. Brooks, S. Robinson, *Simulation*, PALGRAVE, England, Hampshire, 2001.
[11] M. Luck, P. McBurney, Ch. Preist, *Agent technology: Enabling next generation computing: A roadmap for agent-based computing*, AgentLink, 2003.
[12] L. Padgham, M. Winikoff, Prometheus: A Methodology for Developing Intelligent Agents, *Proceedings of the First Intemational Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2002)*, Bologna, Italy, 2002.
[13] F. Parandoosh, Evaluating Agent-Oriented Software Engineering Methodologies, *2007 2nd International Workshop on Soft Computing Applications*, IEEE, 2007.
[14] A. S. Rao, AgentSpeak(L): BDI Agents speak out in a logical computable language, *Proceedings of the Seventh Workshop on Modelling Autonomous Agents in a Multi-Agent World*, 1996.
[15] N. Fornara, M. Colombetti, *A commitment based approach to agent communication*, Applied Artificial Intelligence, Taylor & Francis, 2004.
[16] L. Amgoud, F. Dupin de Saint-Cyr, *A new semantics for ACL based on commitments and penalties*, International Journal of Intelligent Systems, John Wiley and Sons Ltd., 2008.
[17] R. Kibble, Speech acts, commitment and multi-agent communication, *Computational and Mathematical Organization Theory*, 2006.

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-193

Application of Swarm Intelligence in E-Learning Systems

Elżbieta KUKLA
Technical University of Wrocław, Institute for Applied Informatics
Wybrzeże Wyspiańskiego 27, 53-370 Wrocław, Poland
e-mail: elzbieta.kukla@pwr.wroc.pl

Abstract. E-learning systems are intended to offer their users didactic material tailored to their actual needs, preferences and abilities. This idea is consistent with the general principles of constructivism. Additionally it offers a student great flexibility in organizing his own curriculum and his learning scenario in the scope of particular course. However, it has also some side effects causing that students have many problems with making curriculum adapted to their needs or choosing subsequent course elements while complying the course. Many of e-learning courses designers, teachers and students emphasize the necessity of guidance in effective traversing a "knowledge network". The chapter presented below explores some of the recent solutions in this subject. They were carefully selected to represent the widest possible range of research concerning the application of swarm intelligence to recommendation in e-learning systems.

Keywords. e-learning, swarm intelligence, Ant Colony Optimization (ACO) algorithm, adaptive recommendation, curriculum construction, learning scenario determination

Introduction

As the life long learning becomes an obvious necessity for modern society members' technology, enhanced learning becomes more and more popular. This form of learning guarantees the students great flexibility with respect to the time (student learn when he wants and as long as he needs) and place (learning material is available in nearly any place via different medium of communication). Flexibility of e-learning also means great discretion in constructing curriculum and learning scenarios. This often leads student to certain discrepancy. On one hand, it is desirable to have a possibility of choosing what to learn and in which order but on the other hand it demands at least overall knowledge about a learning content, its level of advance and form of presentation. Looking for the solution of these problems students often take advantage of the curriculum or learning scenario suggested by teachers or they ask experienced students (i.e. those who have completed the course or the studies with success).

This problem has been widely investigated by the researches in the domains of: Intelligent Tutoring Systems, Hypertext and Hypermedia Systems, Adaptive Systems, Recommender Systems. All these investigations have delivered very interesting results that will certainly influence future solutions.

The most current direction of the research in this subject is connected with Nature Inspired Systems, namely Swarm Intelligence. The subsequent parts of this chapter explore wide spectrum of these investigations and are supposed to give a representative review of the solutions.

1. Learning environment – general considerations

Modern e-learning environment provides the students with learning material tailored to their needs, preferences and abilities. A student is perceived as an active element of the environment. He can construct his own curriculum and a learning scenario for passing through a particular course. It is possible mainly because of modular construction of the environment based on elementary units called learning objects.

From logical point of view, learning object is a part of knowledge that forms closed and consistent body. Learning object may be of different granularity. It may refer to a course as a whole as well as to a single lesson or even conception or task. Physically learning object may be represented as a text document, presentation, one or more WWW pages, audio or video etc. Note that the same part of knowledge may have different physical representations.

Each learning object is related to certain learning goal that is to be achieved, when a student acquires a part of knowledge represented by the object. Defining learning objects facilitates developing tests that are applied to verify if and how much a student has learnt. Learning goals help with arranging the objects to construct a sequence that permits to pass through learning material. This sequence is often called a learning scenario. The authors of e-learning courses always predefine certain scenarios. For novice and less advanced students it is a very good solution. Nevertheless, more advanced learners, who have already some knowledge and experience in the domain, need an individual approach. They prefer to determine their scenarios tailored to their needs and abilities. Therefore finding an optimal learning scenario for a given student is the subject of scientific research for some years. There were many different proposals for solving this problem. Some of them were based on consensus theory [18] and dynamic classification [19]. Recent investigations concern the application of swarm intelligence for this purpose.

2. Swarm Intelligence - the Foundations

For the first time the expression "swarm intelligence" was used by Beni, Hackwood, and Wang ([4], [5], [6], [7], [8], [9]) with reference to cellular robotic systems. Bonabeau, Dorigo and Theraulaz [2] have extended this definition. By "swarm intelligence" they mean, "any attempt to design algorithms or distributed problem-solving devices inspired by the collective behaviour of social insect colonies and other animal societies" [2].

A group of nature inspired algorithms patterns upon the collective foraging behaviour of ants. They have been applied to combinatorial optimization problems such as Travelling Salesman Problem, graph colouring, sequential ordering etc. This group of algorithms is known as ant-based algorithms or ant colony optimization (ACO) algorithms.

Although they all differ in details, their main idea is the same. It relies on finding an efficient (i.e. short enough) route between nest and food source. This path is not predefined but it emerges because of ants' self-organization that manifests by indirect communication between members of ant colony. In looking for food source an ant is guided by chemical substance (called pheromone) deposited by the other ants. Every time when ant has to select one among many different ways it makes probabilistic choice depending on the amount of pheromone located on each of the alternative branches. At the beginning when there is no pheromone on the branches, each of them can be chosen with equal probability. Nevertheless, if one way leads to food faster than the others, this way will have the greater probability for subsequent ants. This is because the ant returning to nest deposits certain amount of pheromones earlier than the other ants (which have chosen different paths). In this way, shorter solution is reinforced.

One of the simplest models of ant behaviour is based on experiment conducted by Deneubourg et al. [3]. The experiment relied on the investigation how ants choose the way from nest to food source when they are presented with two paths (A and B) of equal length. At the beginning, when none of paths has been visited by any ant, there is no pheromone on any of them. It means that an ant can select each of the two alternative ways with equal probability. As the succeeding ants follow the paths, certain random fluctuations arise. Therefore, more ants than the other visit one of the ways (e.g. A). Since ants following the paths deposit pheromone on it, the way A with more ants have a greater amount of pheromone. Basing on this "information" the succeeding ants will choose the way A. This kind of social communication through the medium of environment is called stigmergy, it places a crucial role in natural, and artificial swarm intelligence based systems.

Supposing that the amount of pheromone deposited on a path is proportional to the number of ants that have selected this way, the probability P_A of choosing way A by next $(i+1)$th ant is:

$$P_A = \frac{(k + A_i)^n}{(k + A_i)^n + (k + B_i)^n} = 1 - P_B \tag{1}$$

where:

A_i, B_i – number of ants that have chosen path A and B accordingly,

k – quantifies the degree of attraction of an unmarked way; the greater k the greater the amount of pheromone to make the choice non-random [2],

n – determines the degree of nonlinearity of the choice function; when the amount of pheromone on one of the ways is inconsiderably greater, great value of n causes that probability of choosing this way by successive ant becomes significantly greater.

In the experiment described in [3] the values of parameters $n \approx 2$ and $k \approx 20$ resulted in the best fitness of model to experimental ants' behaviour. Moreover, if $A_i \gg B_i$ and $A_i \gg 20$, then $P_A \approx 1$ but if $A_i \gg B_i$ and $A_i < 20$, then $P_A \approx 0.5$. The number of ants choosing path A changes dynamically according to P_A value and is:

$$A_{i+1} = \begin{cases} A_{i+1} & \text{if } \delta \le P_A \\ A_i & \text{if } \delta > P_A \end{cases} \tag{2}$$

Similarly, the number of ants choosing path B is computed:

$$B_{i+1} = \begin{cases} B_{i+1} & \text{if } \delta > P_A \\ B_i & \text{if } \delta \le P_A \end{cases} \tag{3}$$

Where δ is a random variable uniformly distributed over [0, 1]. Note that the basic model presented above does not include pheromone evaporation that takes place when observing ant foraging during some period. Nevertheless, it is an excellent explanation and illustration of ants' self-organization.

Other, more complex models take into account time factor and update probability formula of making particular decision. These models are strongly influenced by the problem they solve. Some of them inspired new solutions in e-learning systems presented below in this chapter.

3. ACO Algorithm for Curriculum Construction

Van den Berg et al. [1] have presented an interesting application of ACO algorithms has been presented by The authors focused on the problem of individual curriculum construction. Such an approach offers students flexible choice of most interesting courses for them. The main limitation here is the minimal number of points (i.e. ECTS points) that a student has to attain during his study and sometimes prerequisites that force partial ordering of selected material.

The flipside of the choice freedom is an increasing feeling of loss experienced by students. They have difficulties with establishing the best sequence of courses they will study because they have no experience in the domain as well as it is impossible for them to overview all available courses. Therefore, students looking for the best solution often take advantage of their older colleagues' experiences. This is why swarm intelligence seems to provide a good solution for this problem.

Looking for the solution of the problem mentioned above, the authors [1] have formulated the following assumptions:

- Every learner that takes up a study has his own learning goal described by the level of competence he wants to achieve (e.g. the masters level in a given discipline);
- Each learning goal is associated with certain route, which represents a plan to reach this goal, (it may be curriculum proposed by teachers); the route contains all the courses that should be completed to reach the goal;
- While a student progresses in learning, he constructs his own track consisting of the sequence of learning entities (courses) successfully completed;
- At every decisive moment, (i.e. when a student has to choose subsequent course) his position on the route is determined as his learning track (a set of courses he has completed) extended by the entities that are supposed to be completed.

ACO algorithm for curriculum construction starts with the determination of learning goal by a student. As it was mentioned above, the learning goal involves a route (i.e. a set of courses) leading to reach the goal. When the student faces the decisive problem, the algorithm determines his position on the route. Next, a list of courses that student has to complete is created as the difference between the route and the current student's position. ACO algorithm builds a recommendation for the student by applying transition matrix. This matrix is constructed on the base of the positive experiences of the learners who have completed recommended course starting from the student's position.

Table 1. An exemplary transition matrix [1]

	A	B	C	D	E
A		4	2	5	1
B	2		2	1	3
C	3	4		1	2
D	4	2	4		5
E	1	2	5	3	

An exemplary transition matrix is presented in Table 1. Letters A, B, C, D and E that label rows and columns of matrix, represent a route of a student. Assuming that the student has just completed a course A (it is his position), a list of courses that should be completed next is created. In this case, it comprises courses: B, C, D and E. An appropriate row in transition matrix (labelled “A”) shows how many previous students, starting from the same position A (i.e. A has been completed), have successfully passed courses B, C, D, E. Basing on this information the ACO algorithm recommends the choice of course D as the most often chosen and leading to success. This choice has also the greatest probability of being selected. To prevent sub-optimal convergence to this one solution, the authors introduced certain element of randomness that gives a chance of being selected for the other courses (B, C and E). Final course recommended to the student is being drawn from a set of all possible courses available at the current student's position.

Presented algorithm has been experimentally implemented and (as the authors reported [1]) the obtained results were very promising.

4. Determination of Learning Scenario Using ACO Algorithm

Determining optimal learning scenario for a given student is another field where ACO algorithms have been used. In one of the most often cited works [12] the authors adapted the swarm-based algorithm originally developed to solve the Travelling Salesman Problem [2].

A structure of e-learning material is represented here by a graph. Its nodes refer to “pedagogical items” (exercises, lessons, quizzes etc.) and arcs represent hypertext links between nodes. Moreover, the arcs are labelled by certain weights, which reflect the probabilities that subsequent nodes will be recommended to students. The problem that ACO algorithm has to solve could be formulated as follows: “optimize the weights on the arcs to maximize student's success” [12].

In comparison to the original method, there are some important modifications. Initially pedagogical teams determine weights of graph. It means that at the beginning

not all possible solutions have equal probabilities. Additionally, the authors have assumed two kinds of pheromones: one for success (S) and one for failure (F). Pheromones are deposited backwards. Starting from the last nodes completed by a student and moving back along the path he traversed, an ant (representing the student) deposits decreasing amount of pheromone. The scope of this back propagation has been limited to 4 nodes. Moreover, pheromone evaporates in the course of time. If an arc is not used for a long time, the amount of pheromone tends to 0.

Another new conception introduced in this method is so called historic weight *H* assigned to each node of the graph and every student. Initial *H* value is always equal to 1.0 and it means that the nodes have not been visited yet. If the node has been visited *H* value is modified depending on situation. If a student visiting the node completed his task with a success *H* is multiplied by h_1, otherwise it is multiplied by h_2. Historic weight value is further used to discourage a student from visiting the node he has already seen, especially if visiting the node ended by a success. Since a student has limited memory, *H* value tends to 1.0 in the course of time according to the equation (4):

$$H_t = H_{t-1}\left(1+\frac{1-H_{t-1}}{H_{t-1}}\cdot\frac{1-e^{-\tau x}}{1+e^{-\tau x}}\right) \tag{4}$$

Where x is the time elapsed from the last visit. A constant τ represents the speed of pheromone evaporation.

$$\Leftrightarrow \tau = \frac{1}{x}\cdot\ln\left(\frac{1+\alpha}{1-\alpha}\right)\text{with } \alpha = \frac{H_t - H_{t-1}}{1-H_{t-1}} \tag{5}$$

Since τ value should be calibrated to correspond to student's memory volatility, Eq. (5) [12] can be used for tuning it provided that forgetting time x is precisely determined.

Finally each arc *a* is assigned a "fitness value":

$$f(a) = H\cdot(\omega_1 W + \omega_2 S - \omega_3 F) \tag{6}$$

This value takes into account all the leavens and deterrents for a given node. Fitness value is great if:

- the arc leading to the node was visited long time ago ($H \approx 1.0$);
- the arc is recommended by teachers' team (*W* is high);
- there are many students who have succeeded choosing this node (high S);
- there are not many learners who failed selecting this node (low F).

The higher is $f(a)$ the more likely the node will be recommended for students [13]. It is worth to notice that the contribution of *H*, *W*, *S* and *F* factors in final $f(a)$ value is adjusted by the ω_1, ω_2, ω_3 coefficients.

Fitness value is then used to arrange all the considered nodes that have been evaluated in previous steps. This sequence is a basic list for making a choice.

The authors notice [13] that the choice of an appropriate selection mechanism is not a trivial task because it influences:

- speed with which an arc that emerges as excellent (according to its fitness value) reaches the state of being the only solution recommended to students;
- information loss that may occur when selection method is based only on the ranking created according to the fitness value;
- necessity of very careful and time-consuming calibration of algorithm parameters.

In this context five selection mechanisms previously used in the domain of Genetic Algorithms have been tested. All of them provide different rules for probability determination and make different selection based on counted values and probability distributions. Finally, a stochastic tournament selection method has been applied as the method that accomplishes the best the conditions mentioned above.

5. Swarm Intelligence Algorithms Extended by the Learning Styles Models

An interesting approach to the application of swarm intelligence in e-learning systems has been presented by Wang et al. [11]. The authors have been supposed that students' community is not homogeneous. Its members differ with respect of many features from which an individual learning styles is one of the most important. Hence, the idea of extending adaptive recommendation system based on ACO algorithm with learning style conception.

For extension a relatively simple learning styles model *VARK* has been used. Subsequent letters in *VARK* abbreviation means accordingly: *V* – Visual, *A* – Aural/Auditory, *R* – Read/Write, *K* – Kinaesthetic sensory modalities that are used for learning [20]. Establishing such a solution it is necessary to make additional assumptions. Firstly, a learning style of a given student is known, i.e. the student declares his learning style or he/she fills in a questionnaire that is supposed to determine student's learning style. Secondly, a learning content (or learning objects) are described in categories of *VARK* modalities. These assumptions led the authors to the concept of "homogeneous ant" [11] that is a group of learners that evince nearly the same behaviours in a given learning situation. This group seems to be well described by the learning style mentioned above. Formally, "homogeneous ant" is defined as follows:

Let k be an ant with one of the *VARK* learning styles. When k visits a node it deposits a fixed amount of pheromone. By "homogeneous ant" it is meant another ant that in the same node has the same learning style as k. Furthermore, if the node visited by ant k is described as matching to k learning style it may obtain an extra amount of pheromone.

An adaptive rule applied in this system (called by the authors SACS – Style-based Ant Colony System) is well explained by the following example. When a learner has complied a learning content of a node a and he wants to continue his study, the system examines a learning transition matrix. Like in previously presented swarm intelligence applications, learning transition matrix contains the numbers of students that have passed from a node denoted by row label to a node described by column label. The matrix analysis provides a set of nodes that may be visited next. Then, an additional procedure (called adaptive rule) reanalysis the result obtained to determine nodes that the best suits to student's learning style, i.e. the procedure will recommend the nodes which have been visited by more homogeneous ants. To say exactly these nodes will

have the higher probabilities to be recommended since the amount of pheromone deposited on the paths leading to them is higher than in case of the remaining nodes.

Besides modifications presented above ACO algorithm applied in SACS uses a learning forgetting curve to incorporate time factor that in turn is important with regard to the pheromone evaporation. The forgetting function is used to determine the ratio of learning memory retention i.e. a learner's ability necessary to achieve the next learning goal. The ratio of learning memory retention constitutes the heuristic information for the system.

Finally, the algorithm for finding adaptive path consists of six following steps [11]:

1. Parameter initialization – this step is treated as pre-processing and aims at determining initial values of parameters used in further calculations; at this step every student is assigned a number referring to his learning style: 1 – Visual, 2 – Aural/Auditory, 3 – Read/Write, 4 – Kinaesthetic.
2. The learning transition matrix creation – for a given time *t* learning transition matrix is created upon the information gathered by system in previous moment *t*-1.
3. Computing heuristic information – the information is calculated as a ratio of learning memory retention for every node.
4. Updating the pheromone trails – updating rule takes into account total number of number of ants that have completed learning material contained in the node, number of homogeneous ants and pheromone evaporation.
5. Counting the probability distribution – for each node a probability of being recommended to a student is calculated.
6. Listing the recommended paths – the recommended paths are chosen basing on the probability distribution calculated in the preceding step.

Y. J. Yang and C. Wu [10] have proposed slightly different solution. The authors propose using Kolb's learning style inventory [21] - more complex model then previously used. The assumptions that have lied at the bottom of presented solution are very similar to those presented for SACS [11]. His competence level and his learning style characterize every student. In addition, each learning object is associated with two attributes: learning object type and learning object level. The attributes values are supposed to be annotated by teachers or learning content providers and they are known to system when the student begins learning.

The difference between two presented systems is an ant interpretation. In [10] Attribute-based Ant Colony System (AACS) an idea of an "attribute ant" has been introduced. The attribute ant uses an extended learner model (comprising his learning style and competence level) as well as object level to run an adaptive procedure for selecting the best learning object. The object selection results from a rule of the pheromone updating. In the AACS there are two rules [10]:

1. **IF** *learner's attributes* **MATCH** *the learning object's attributes* **THEN** *update pheromone trials with* **MR**(*style*) **AND** **MR**(*level*).
2. **IF** *learner's attributes* **PARTIALLY MATCH** *the learning object's attributes* **THEN** *update pheromone trials with* **MR**(*style*) **OR** **MR**(*level*).

The meanings of the conditions MATCH and PARTIALLY MATCH have been defined as follows:

1. *Learner's attributes* **MATCH** *learning object's attributes* **IF** *learner's learning style = learning object type* **AND** *learner's domain knowledge level = learning object's level.*

2. *Learner's attributes* **PARTIALLY MATCH** *learning object's attributes* **IF** {*learner's learning style = learning object type* **AND** *learner's domain knowledge level* ⟨⟩ *learning object's level*} **OR** {{*learner's learning style* ⟨⟩ *learning object type* **AND** *learner's domain knowledge level = learning object's level*}.

MR(style) and MR(level) denote a matching ratio for a learning object with respect to student's learning style and his domain knowledge level. Learner's and learning object's attributes adjustment is presented in Table 2.

Table 2. The learner's and learning object's attributes adjustment [10]

Learner's attributes		Learning object's attributes	
Kolb's Learning Style	Knowledge level	Learning object type	Learning object level
Diverging	Apprentice	Graphic (image, chart)	Initial
Assimilating	Beginner	Video (audio, animation)	Introductory
Converging	Intermediate	Text (word, PowerPoint)	Advance
Accommodating	Expert	XML (web, SCORM, LOM)	Professional

Summarizing, the heuristic information can be reinforced if an attribute ant matches or partially matches the attributes of a learning object. In this case an extra amount of pheromone is deposited according to the MR(style) and MR(level) values.

The procedure of adaptive way finding is very similar to the algorithm presented above for SACS and it consists of the following steps:

1. Parameter initialization,
2. Constructing solution (basing on transition probability values),
3. Determination of the heuristic decision rule,
4. Updating pheromone trails,
5. Global pheromone updating taking into account attribute ants,
6. Learning objects recommendation.

Presented system has been implemented as an intelligent agent working in web environment. The agent sequences a query answers according to user's characteristics that comprises student's learning style and his advance in knowledge domain.

6. Final Remarks

As it was mentioned in the introduction for the chapter, the application of swarm intelligence in e-learning systems is relatively new direction of investigations. The solutions presented above are different modifications of basic ACO algorithm originally developed for solving Travelling Salesman Problem (TSP).

They all make implicit or explicit assumptions that may but need not be fulfilled in reality. One of the implicit assumptions is a structure of graph. In TSP a graph represents cities (nodes) connected by routes (arcs). In e-learning environment, a graph represents learning objects (nodes) and relations between objects (nodes). What is the discrepancy? In TSP an optimization task is defined as "finding the shortest Hamiltonian way in finite graph". In the e-learning applications, a graph should not be considered as finite because all the time new learning objects are added. For this reason,

optimization task does not relate to Hamiltonian way (that includes all nodes of the graph). There is also another reason that seems to be important in rejecting Hamiltonian way as a solution for e-learning. In learning network there are often certain redundancy, there are many learning objects that refer to the same content. Each of these objects represents another approach to teaching the same part of learning material, each of them uses different form and media of presentation etc. Constructing curriculum or learning scenario means choosing only one of these objects (not all of them). Therefore, a task for e-learning problem solution is not the same as for TSP. Moreover, the solution of TSP does not include the situation when certain nodes require some other nodes to be visited earlier. This is known as sequential ordering problem [2], which generally relies on finding a path on a directed graph when the nodes and the arcs are assigned certain weights. Moreover, there are precedence constrains that force nodes visiting in partially predefined order. The path found as a solution should be of a minimum weight Hamiltonian way.

In turn, the problem of curriculum construction when the only restriction is the total amount of points that should be gained during the studies is very similar to the problem commonly known as "knapsack problem" that relies on knapsack repletion with a luggage of a given volume. Knapsack has its own volume and the task is accomplished when knapsack volume is filled. It worth to notice that if volume means ECTS points the curriculum constructed in this way fulfils formal demands but has nothing to do with learning content and course sequencing.

These initial remarks together with the open questions and problems presented in [14] and [15] should be the foundations and inspiration for further investigations in this domain.

References

[1] B. van den Berg, C. Tattarsal, J. Jansses, F. Brouns, H. Kurvers, R. Koper, Swarm-based Sequencing Recommendations in E-learning, *International Journal of Computer Science & Applications* 3 (2006), 1-11.

[2] E. Bonabeau, M. Dorigo, G. Theraulaz, *Swarm Intelligence. From natural to Artificial Systems*, Oxford University Press, New York, Oxford, 1999.

[3] J.-L. Deneubourg, S. Aron, S. Goss, J.-L.Pasteels, The self-Organizing Exploratory Pattern of the Argentine Ant, *Journal of Insect Behaviour* 3 (1990), 159-168.

[4] G. Beni, The concept of Cellular Robotic system, *Proceedings of IEEE International Symposium on Intelligent Control*, Los Alamitos, CA: IEEE Computer Society Press, 1988, 57-62.

[5] G. Beni, J. Wang, Swarm Intelligence. *Proceedings of Seventh Annual Meeting of the Robotic Society of Japan*, RSJ Press, Tokyo, 1989, 425-428.

[6] G. Beni, J. Wang, Theoretical Problems for the Realization of Distributed Robotic Systems, *Proceedings of the IEEE International Conference on Robotic and Automation*, Los Alamitos, CA: IEEE Computer Society Press, 1991, 1914-1920.

[7] G. Beni, S. Hackwood, Stationary Waves in Cyclic Swarms, *Proceedings of IEEE International Symposium on Intelligent Control*, Los Alamitos, CA: IEEE Computer society Press, 1992.

[8] S. Hackwood, G.Beni, Self-Organizing Sensors by Deterministic Annealing. *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, IROS'91, Los Alamitos, CA: IEEE Computer Society Press, 1991, 1177-1183.

[9] S. Hackwood, G. Beni, Self-Organization of Sensors for Swarm Intelligence, *Proceedings of the International Conference on Robotics and Automation*, Los Alamitos, CA: IEEE Computer Society Press, 1992, 819-829.

[10] Y. J. Yang, C. Wu, An attribute-based ant colony system for adaptive learning object recommendation, *Expert Systems with Applications* (2008), doi:10.1016/j.eswa2008.01.066.

[11] T.-I. Wang, K.-T. Wang, Y.-M. Huang, Using a style-based ant colony system for adaptive learning. *Expert systems with Applications* 34 (2008), 2449-2464.

[12] Y. Semet, Y. Jamont, R. Biojout, E. Lutton, P. Collet, *Artificial Ant Colonies and E-learning : An Optimization of Pedagogical Paths*.

[13] Y. Semet, E. Lutton, P. Collet, Ant Colony Optimisation for E-learning: Observing the Emergence of Pedagogic Suggestions, *IEEE*, 2003.

[14] Z. Chen, Learning about Learners: System Learning in Virtual Learning Environment, *International Journal of Computers, Communications & Control* III (2008), 33-40.

[15] C. Tattersall, J. Janssen, B, van den Berg, R. Koper, *Social Insect-inspired e-Learning: Open Research Questions*.

[16] S. Guitierrez, A. Pardo, C. D. Kloos, *Finding a learning path: Towards a swarm intelligence approach*.

[17] S. Gutierrez, G. Valigiani, P. Collet, C. D. Kloos, *Adaptation of the ACO heuristics for sequencing learning activities*.

[18] E. Kukla, N. T. Nguyenh, C. Daniłowicz J. Sobecki, M. Lenar, A model conception for optimal scenario determination in a intelligent learning system. *Interactive Technology and Smart Education* 1 (2004), 171-183.

[19] C. Daniłowicz, E. Kukla, The application of adaptive students' classification to the determination of a learning strategy in an e-learning environment. *World Transactions on Engineering and Technology Education* 2 (2003), 395-398.

[20] N. Fleming, *VARK - A guide to learning styles*, http://www.vark-learn.com.

[21] D. A. Kolb., *Experimental learning: Experience as the source of learning and development*, Prentice Hall, New Jersey 1984.

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-204

Determination of Opening Learning Scenarios in Intelligent Tutoring Systems

Adrianna KOZIERKIEWICZ
Institute of Information Science and Engineering, Wrocław University of Technology
e-mail: adrianna.kozierkiewicz@pwr.wroc.pl

Abstract. The main purpose of intelligent tutoring systems is to guarantee an effective learning and offer the optimal learning path for each student. Therefore, determination of learning scenario is a very important task in a learning process. After a new student is registered in the system, he is classified to the appropriate group. Before he begins to learn an opening scenario is determined based on final scenarios of students who belong to the class of learners similar to the new one. The new student is offered the optimal learning path suitable for his preferences, learning styles and personal features. In this paper new knowledge structure, which involves version of lessons, is proposed. For the defined knowledge structure definitions of learning scenario, distance function and the procedure of the scenario determination are presented.

Keywords. intelligent tutoring system, knowledge structure, learning scenario

Introduction

In the recent years we can observe new trend in education: traditional learning is supported by e-learning or even learning with using new technology such as CDs, MP3 Players, computers, mobile phones and computer network replaces traditional learning with teachers and classrooms. This fact is characterized by the speed of scientific and technological progress which implicates need of lifelong learning. People want to improve their qualifications, bring their skills up-to-date and retrain a new line of work and demand the guarantee of effective learning in a short and flexible time. This requirement could be fulfilled by using intelligent tutoring systems.

In the proposed system the learning concept is based on an assumption that similar students will learn in the same or a very similar way [5]. The learning process consists of several steps. Before starting the learning process students provide information about themselves. It is a very important task because the intelligent tutoring system offers the optimal learning path suitable for each student. Based on demographic data, learning styles, abilities, personal character traits, interests and current levels of knowledge [4] system proposes the most adequate educational material in the best order. After registration process student is classified to a group of similar learners. For that learner an opening scenario is determined based on scenarios of learners who belong to the same class. Student is presented with a sequence of versions of lessons corresponding to one unit. When he finishes he has to pass a test. System assesses student's results and tries to find the reason of mistakes. The test score determines the next step. If system decides that learner achieves sufficient score he can proceed to the

next sequence relating to the next unit. Otherwise, student is suggested relearning by using different presentation methods. The opening scenario is being modified during the learning process. The final scenario is stored in the user's profile. Sometimes, changes could cause the student to be reclassified to a different group. Figure 1 presents the described learning process.

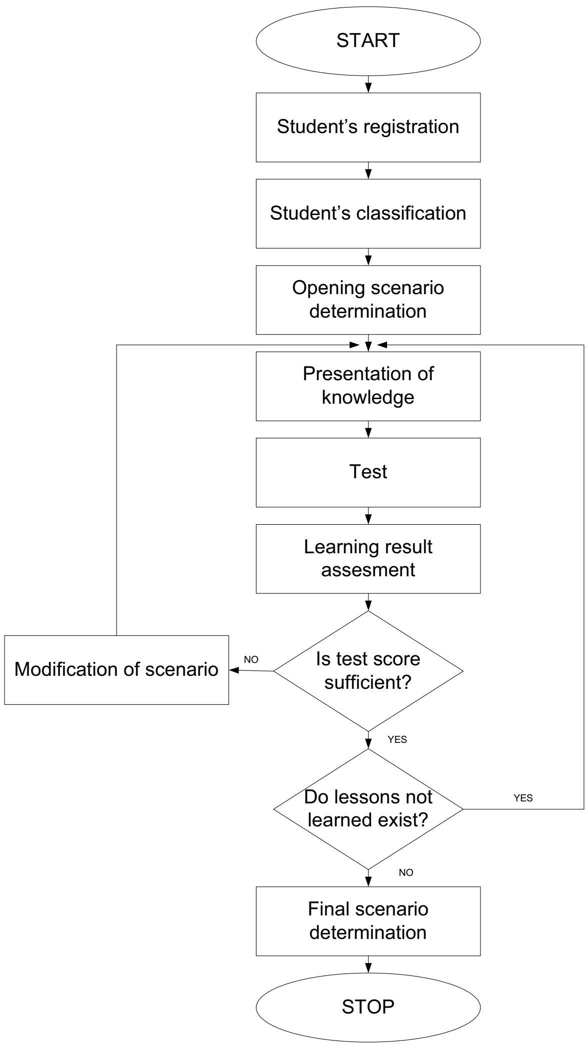


Figure 1. Learning process [5]

This chapter is devoted to the method of determining the learning scenario for a student who starts to learn. The opening scenario is generated based on final scenarios of students who belong to the same class. Such solution personalizes the learning environment and offers the optimal learning path suitable for each student.

A new knowledge structure and a learning scenario are defined. Knowledge structure described in [7] is extended with lessons versioning. The proposed knowledge

structure is more flexible and allows adapting suitable version of lesson for student's preferences. A very important task is measuring differences between two given scenarios. Distance function includes three elements: distance between units and their order, distance between lessons and their order proposed in [7] and distance between versions of lessons. Finally, algorithm for determining opening learning scenario based on consensus theory is worked out. To the procedure described in [7] a new, third part is added which is responsible for choosing versions of lessons.

In the next section methods used in e-learning systems to curriculum sequencing are presented. Subsection 3 contains definitions of knowledge structure and a learning scenario. The procedure of the scenario determination with necessary definitions is presented in section 4. Finally, conclusion and future works are introduced.

1. Related Works

Curriculum sequencing is the oldest and most popular technology applied in intelligent and adaptive e-learning systems. There are two different kinds of sequencing: active and passive. First of them suggests a learning goal (a subsets of concepts or topics). System with active sequencing can build the optimal and an individual learning path to achieve the assumed goal. Passive sequencing begins when student is not able to solve a problem or answer a question correctly. System offers then an available learning material which can fill the missing of knowledge. In intelligent tutoring systems there are two levels of sequencing: high and low. High-level sequencing refers to orders of concepts, topics and lessons. Low-level determines tasks, problems, tests and examples [1], [2].

In ELM-ART sequencing is implemented in a form of a recommend link and an adaptive „next" button. System provides special links in the form of coloured bullets which allow you to navigate between units of knowledge. The colour of this bullet informs the student about educational status of next pages [8].

Active Math offers dynamically constructed courses. In the process of course generation different kinds of information such as: goal, scenario the user choose, level of knowledge, interaction history with system, student's capabilities, pedagogical rules are taken to determine when and which items should be presented and in which order. It includes the four steps course generator. The concepts which need to be learned by students are chosen in the first step. The next step let one choose all additional items: examples, exercises and tests. In the third step, pedagogical rules and information from the user's model are used to select an appropriate content. Finally, the pages are ordered and put together in a proper hierarchy [6].

In INSPIRE the Lesson Generation Module is responsible for planning the learning path. In a lesson generating process one of 3 levels (remember, use, find) can be chosen to present outcome concepts. The generated lesson includes: a set of presentations of the outcome concepts, links to brief presentations of the prerequisite concepts focusing on their relation with the outcome and links to definitions of related concepts. Lessons are generated based on student's knowledge level and learning styles [3].

Curriculum sequencing is also applied in: KBS Hyperbook, InterBook, PAT InterBook, CALAT, VC Prolog Tutor, ELM-ART-II, AST, ADI, ART-Web, ACE and ILESA. In [5] a method of determining a learning scenario using consensus methods is proposed.

2. Knowledge Structure

In intelligent E-learning systems the educational material should not only have a textual form but also should include multimedia materials and be presented suitable for student's preferences, learning styles and personal features. Domain knowledge is represented in three levels: the first one is logical which refers to units and relations between them, the second one is multimedia lessons and relations between them and the last is related to other forms of lesson for example: textual, graphical, interactive etc. Let $C = \{c_1,...,c_q\}$ be the finite set of units that is elementary, indivisible part of knowledge. By P_i we denote the set of lessons corresponding to unit c_i where $i \in \{1,...,q\}$, $P = \bigcup_{i=1,...,q} P_i$. Each lesson p_i is related to their version $v^{(i)}{}_{jk}$ for $k \in \{1,...,m\}$, $V = \bigcup_{i=1,...,q} \bigcup_{j=1,..,r_i} \bigcup_{k=1,...,m} v^{(i)}{}_{jk}$, r_i=*card*(P_i). *Rc* is called a set of linear relations between units. Each of such relations defines the order in which the modules should be presented to a student, because some units should be learned before others. A binary relation γ is called linear if relation is reflexive, transitive, antisymmetric and total. By $Rp(P_i)$ we denote a set of partial linear relations between lessons where $i \in \{1,...,q\}$, $Rp = \bigcup_{i=1,...,q} Rp(P_i)$. It means that some lessons should be presented before others. A binary relation γ is called partial linear if relation is reflexive and antisymmetric.

The proposed knowledge structure makes the choice of units' and lessons' order and versions of lessons possible. The graphic representation of defined knowledge structure is presented in the Figure 2.

3. Learning Scenario Definition

For knowledge structure described in section 3 a new definition of a learning scenario is introduced. This definition contains information about units, lessons and version of lessons but does not contain information about tests. It is defined in the following way:

Definiton 1. *By a learning scenario s based on orders* $(\alpha, \beta_1, \beta_2,..,\beta_q)$ *we call a sequence of versions of lessons:* $s = < w_1, w_2,..., w_q >$ *where sequence* w_i *refers to exactly one unit* c_{ρ_i} *and s does not contain any other sequence referring to the unit* c_{ρ_i}, *for* $i, \rho_i \in \{1,...,q\}$, ρ_i *- number of unit referring to a sequence of versions of lessons* w_i, $\alpha \in Rc$, $\beta_i \in Rp$,

A learning scenario fulfils the following conditions:

1. *The order of any 2 sequences* w_i, w_j *in s should correspond to the order of their units in relation* $(c_{\rho_i}, c_{\rho_j}) \in \alpha$, *for* $i \neq j$ *and fixed* $i, j, \rho_i, \rho_j \in \{1,...,q\}$.

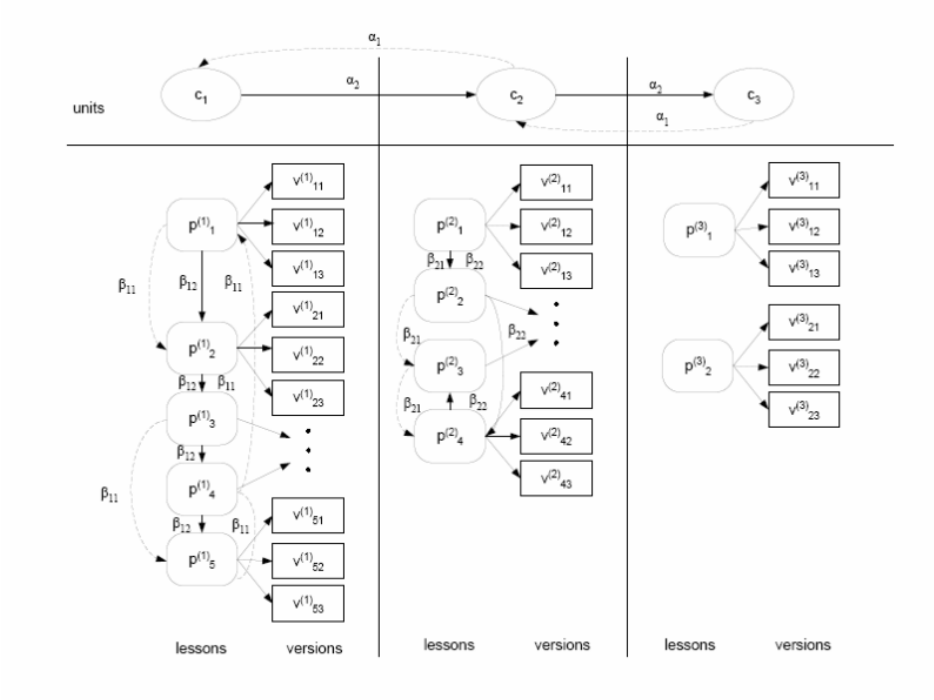


Figure 2. Knowledge structure

2. *The order of any versions of lessons in a sequence* w_i *should correspond to the order in relation* β_{ρ_i} , $w_i \in \beta_{\rho_i}$ *for fixed* $i, \rho_i \in \{1,...,q\}$.
3. *If* s *contains* $v^{(i)}{}_{jk}$, s *does not contain* $v^{(i)}{}_{xy}$ *where* $x = j$, $y = k$ *for fixed* $i \in \{1,...,q\}$, $j, x \in \{1,...,r_i\}$, $k, y \in \{1,...,m\}$.

The following example should help with understanding the definition.

Example 1.

Let consider knowledge structure presented in Figure 1. Some scenarios are defined as follows:

$$s^{(1)} = << v^{(1)}{}_{11}, v^{(1)}{}_{21}, v^{(1)}{}_{31}, v^{(1)}{}_{42}, v^{(1)}{}_{52} >, < v^{(2)}{}_{12}, v^{(2)}{}_{22}, v^{(2)}{}_{33}, v^{(2)}{}_{43} >, < v^{(3)}{}_{22} >>$$

$$s^{(2)} = << v^{(1)}{}_{31}, v^{(1)}{}_{52}, v^{(1)}{}_{42}, v^{(1)}{}_{11}, v^{(1)}{}_{21} >, < v^{(2)}{}_{12}, v^{(2)}{}_{22}, v^{(2)}{}_{42}, v^{(2)}{}_{32} >, < v^{(3)}{}_{21} >>$$

$$s^{(3)} = << v^{(3)}{}_{21} >, < v^{(2)}{}_{12}, v^{(2)}{}_{22}, v^{(2)}{}_{42}, v^{(2)}{}_{32} >, < v^{(1)}{}_{31}, v^{(1)}{}_{52}, v^{(1)}{}_{42}, v^{(1)}{}_{11}, v^{(1)}{}_{21} >>$$

4. Determination of Learning Scenario

After user’s registration and classification student can start learning process. Before he begins to learn an opening scenario is chosen by the system. The opening scenario is chosen from final scenarios of students who belong to the class of learners similar to the new one.

The proposed knowledge structure enables creating different courses and learning scenarios. Creating learning scenario is conducted in three steps. First of them depends on a proper order of units. Each unit is related to sets of lessons. The choice of suitable lessons' order is the second step of creating learning scenario. In the last step versions of lessons are chosen.

In this section an algorithm based on a consensus theory is presented. The consensus problem is presented as follows: *For given learning scenarios* $s^{(1)}, s^{(2)}, \ldots, s^{(n)}$ *one should determine* s^* *such that the condition* $\sum_{i=1}^{n} d(s^*, s^{(i)}) = \min_{s} \sum_{i=1}^{n} d(s, s^{(i)})$ *is satisfied.* For comparing two different scenarios $s^{(1)}$ and $s^{(2)}$ we define distance function in the following way:

Definition 2. *By* $d: S_C \times S_C \to [0,1]$ *we call a distance function between scenarios* $s^{(1)}$ *and* $s^{(2)}$. *Distance function is calculated as:*

$$d(s^{(1)}, s^{(2)}) = \lambda_1 \sigma(\alpha^{(1)}, \alpha^{(2)}) + \lambda_2 \frac{\sum_{i=1}^{q} \sigma(w^{(1)}{}_{g_i}, w^{(2)}{}_{h_i})}{q} + \lambda_3 \frac{\sum_{i=1}^{q} \delta(w^{(1)}{}_{g_i}, w^{(2)}{}_{h_i})}{q}$$

where S_C *- set of learning scenarios,* $\lambda_1 + \lambda_2 + \lambda_3 = 1$, g_i, h_i *– the position referring to unit* c_i *in scenarios* $s^{(1)}$ *and* $s^{(2)}$, *respectively.*

The value of distance function $d(s^{(1)}, s^{(2)})$ we can estimate in three steps:

1) $\sigma(\alpha^{(1)}, \alpha^{(2)}) = \dfrac{\sum_{i=1}^{q} S(\alpha^{(1)}, \alpha^{(2)})}{q}$

where: $S(\alpha^{(1)}, \alpha^{(2)}) = \dfrac{|k^{(1)} - k^{(2)}|}{q}$ if unit c occurs in $\alpha^{(1)}$ on position $k^{(1)}$ and in $\alpha^{(2)}$ on position $k^{(2)}$.

2) $\sigma(w^{(1)}{}_{g_i}, w^{(2)}{}_{h_i}) = \dfrac{\sum_{i=1}^{r_i} S(w^{(1)}{}_{g_i}, w^{(2)}{}_{h_i})}{r_i}$

where: $S(w^{(1)}{}_{g_i}, w^{(2)}{}_{h_i}) = 0$ if lesson p not occur in either $w^{(1)}{}_{g_i}$ or $w^{(2)}{}_{hi}$;

$S(w^{(1)}{}_{g_i}, w^{(2)}{}_{h_i}) = \dfrac{|k^{(1)} - k^{(2)}|}{r_i}$ if lesson p occurs in $w^{(1)}{}_{g_i}$ on position $k^{(1)}$ and in $w^{(2)}{}_{hi}$ on position $k^{(2)}$; $S(w^{(1)}{}_{g_i}, w^{(2)}{}_{h_i}) = \dfrac{r_i - (k-1)}{r_i}$ if lesson p occurs in $w^{(1)}{}_{g_i}$ on position k and does not occur in $w^{(2)}{}_{hi}$ (or occurs in $w^{(2)}{}_{hi}$ on position k but does not occur in $w^{(1)}{}_{g_i}$).

3) $\delta(w^{(1)}{}_{g_i}, w^{(2)}{}_{h_i}) = \dfrac{\sum_{j=1}^{r_i} \theta(v^{(1)(i)}{}_{jk}, v^{(2)(i)}{}_{jy})}{\max\{card(w^{(1)}{}_{g_i}, w^{(2)}{}_{h_i})\}}$

where: $\theta(v^{(1)(i)}{}_{jk}, v^{(2)(i)}{}_{jy}) = \begin{cases} 1, if\ v^{(1)(i)}{}_{jk} \neq v^{(2)(i)}{}_{jy} \\ 0\ \ otherwise \end{cases}$, for $k, y \in \{1,...,m\}$

The proposed distance function is a metric. It consists of three elements being metrics: for first and second one proof is given in [7], the third is the Hamming metric.

Example 2

Let $s^{(1)} =<<v^{(1)}{}_{11}, v^{(1)}{}_{21}, v^{(1)}{}_{31}, v^{(1)}{}_{42}, v^{(1)}{}_{52}>, <v^{(2)}{}_{12}, v^{(2)}{}_{22}, v^{(2)}{}_{33}, v^{(2)}{}_{43}>, <v^{(3)}{}_{22}>>$

$s^{(2)} =<<v^{(1)}{}_{31}, v^{(1)}{}_{52}, v^{(1)}{}_{42}, v^{(1)}{}_{11}, v^{(1)}{}_{21}>, <v^{(2)}{}_{12}, v^{(2)}{}_{22}, v^{(2)}{}_{42}, v^{(2)}{}_{32}>, <v^{(3)}{}_{21}>>$

and $\lambda_1 = \lambda_2 = \lambda_3 = \dfrac{1}{3}$.

The distance function $d(s^{(1)}, s^{(2)})$ between scenarios $s^{(1)}$ and $s^{(2)}$ is equal:

1) $\sigma(\alpha^{(1)}, \alpha^{(2)}) = \dfrac{\sum_{i=1}^{q} S(\alpha^{(1)}, \alpha^{(2)})}{q} = \dfrac{0}{3} = 0$

2) $\sigma(w^{(1)}{}_1, w^{(2)}{}_1) = \dfrac{\sum_{i=1}^{r_i} S(w^{(1)}{}_1, w^{(2)}{}_1)}{r_1} = \dfrac{12}{25}$

$\sigma(w^{(1)}{}_2, w^{(2)}{}_2) = \dfrac{\sum_{i=1}^{r_i} S(w^{(1)}{}_2, w^{(2)}{}_2)}{r_2} = \dfrac{1}{8}$

$\sigma(w^{(1)}{}_3, w^{(2)}{}_3) = \dfrac{\sum_{i=1}^{r_i} S(w^{(1)}{}_3, w^{(2)}{}_3)}{r_3} = 0$

3) $\delta(w^{(1)}{}_1, w^{(2)}{}_1) = \dfrac{\sum_{j=1}^{r_i} \theta(v^{(1)(1)}{}_{jk}, v^{(2)(1)}{}_{jy})}{\max\{card(w^{(1)}{}_1, w^{(2)}{}_1)\}} = 0$

$\delta(w^{(1)}{}_2, w^{(2)}{}_2) = \dfrac{\sum_{j=1}^{r_i} \theta(v^{(1)(2)}{}_{jk}, v^{(2)(2)}{}_{jy})}{\max\{card(w^{(1)}{}_2, w^{(2)}{}_2)\}} = \dfrac{2}{4}$

$$\delta(w^{(1)}{}_3, w^{(2)}{}_3) = \frac{\sum_{j=1}^{r_i} \theta(v^{(1)(3)}{}_{jk}, v^{(2)(3)}{}_{jy})}{\max\{card(w^{(1)}{}_{3_i}, w^{(2)}{}_3)\}} = \frac{1}{1} = 1$$

$$d(s^{(1)}, s^{(2)}) = \lambda_1 \sigma(\alpha^{(1)}, \alpha^{(2)}) + \lambda_2 \frac{\sum_{i=1}^{q} \sigma(w^{(1)}{}_{g_i}, w^{(2)}{}_{h_i})}{q} + \lambda_3 \frac{\sum_{i=1}^{q} \delta(w^{(1)}{}_{g_i}, w^{(2)}{}_{h_i})}{q} =$$

$$= \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot \frac{\left(\frac{12}{25} + \frac{1}{8}\right)}{3} + \frac{1}{3} \cdot \frac{\left(\frac{1}{2} + 1\right)}{3} = \frac{121}{1800} + \frac{1}{6} = \frac{421}{1800} \approx 0.234$$

The procedure of determining the learning scenario is presented as follows:

Given: $s^{(1)}, s^{(2)}, \ldots, s^{(n)}$

Result: s^*

BEGIN

1. i=1;
2. For c_i determine a set with repetitions:

   $I(c_i) = \{j : \textit{there exists a scenario that } c_i \textit{ occurs on its jth position}\}$
3. Calculate $J(c_i) = \sum_{j \in I(c_i)} j$
4. i++;
5. If $i<q$ GOTO 2.
6. Set sequences of version of lesson in relation α^* according to the increasing of values $J_c(c_i)$;
7. i=1;
8. j=1;
9. $w^* = w^{(j)}{}_{g_i}$, $\Omega = \sum_{k=1}^{n} \sigma(w^*, w^{(k)}{}_{g_i})$;
10. j++;
11. If $\sum_{k=1}^{n} \sigma(w^{(j)}{}_{g_i}, w^{(k)}{}_{g_i}) < \Omega$ then $\Omega = \sum_{k=1}^{n} \sigma(w^{(j)}{}_{g_i}, w^{(k)}{}_{g_i})$; $w^* = w^{(j)}{}_{g_i}$;
12. If $j<n$ GOTO 10.
13. Set lesson's partial order for unit c_i like in w^*
14. i++;
15. If $i<q$ GOTO 8.
16. i=1;
17. j=1;
18. For each $k \in \{1,\ldots,m\}$ calculate $f(v^{(i)}_{jk})$, that is number of appearances $v^{(i)}_{jk}$

in $w_{g_i}{}^{(1)}, w_{h_i}{}^{(2)}, \ldots, w_{d_i}{}^{(n)}$;

19. For lesson $p_j^{(i)}$ choose version of lesson $v_{jt}^{(i)}$ such that $f(v_{jt}^{(i)}) = \max_{k \in \{1,\ldots,m\}} f(v_{jk}^{(i)})$;
20. j++;
21. If $j<n$ GOTO 18
22. i++;
23. If $i<q$ GOTO 17.

END

Example 3.

Let

$s^{(1)} = \langle\langle v^{(1)}{}_{11}, v^{(1)}{}_{21}, v^{(1)}{}_{31}, v^{(1)}{}_{42}, v^{(1)}{}_{52} \rangle, \langle v^{(2)}{}_{12}, v^{(2)}{}_{22}, v^{(2)}{}_{33}, v^{(2)}{}_{43} \rangle, \langle v^{(3)}{}_{22} \rangle\rangle$

$s^{(2)} = \langle\langle v^{(1)}{}_{31}, v^{(1)}{}_{52}, v^{(1)}{}_{42}, v^{(1)}{}_{11}, v^{(1)}{}_{21} \rangle, \langle v^{(2)}{}_{12}, v^{(2)}{}_{22}, v^{(2)}{}_{42}, v^{(2)}{}_{32} \rangle, \langle v^{(3)}{}_{21} \rangle\rangle$

$s^{(3)} = \langle\langle v^{(3)}{}_{21} \rangle, \langle v^{(2)}{}_{12}, v^{(2)}{}_{22}, v^{(2)}{}_{42}, v^{(2)}{}_{32} \rangle, \langle v^{(1)}{}_{31}, v^{(1)}{}_{52}, v^{(1)}{}_{42}, v^{(1)}{}_{11}, v^{(1)}{}_{21} \rangle\rangle$

First we calculate units' order.

We have:

$I(c_1) = \{1,1,3\}$,

$I(c_2) = \{2,2,2\}$,

$I(c_3) = \{3,3,1\}$.

and $J(c_1) = 5$, $J(c_2) = 6$, $J(c_3) = 7$

thus $\alpha^* = \langle c_1, c_2, c_3 \rangle$

Now we have to choose lessons' order.

For c_1:

If $w^* = w^{(1)}{}_1$ then $\Omega = \frac{12}{25} + \frac{12}{25} = \frac{24}{25}$;

If $w^* = w^{(2)}{}_1 = w^{(3)}{}_1$ then $\Omega = \frac{12}{25} + 0 = \frac{12}{25}$; so $w^* = w^{(2)}{}_1 = w^{(3)}{}_1$

For c_2:

If $w^* = w^{(1)}{}_2$ then $\Omega = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$;

If $w^* = w^{(2)}{}_2 = w^{(3)}{}_2$ then $\Omega = \frac{1}{8} + 0 = \frac{1}{8}$; so $w^* = w^{(2)}{}_2 = w^{(3)}{}_2$

For c_3:

If $w^* = w^{(1)}{}_3 = w^{(2)}{}_3 = w^{(3)}{}_3$ then $\Omega = 0$; so $w^* = w^{(1)}{}_3 = w^{(2)}{}_3 = w^{(3)}{}_3$

Finally we select version of lessons:

For c_1 we choose: $v^{(1)}{}_{11}, v^{(1)}{}_{21}, v^{(1)}{}_{31}, v^{(1)}{}_{42}, v^{(1)}{}_{52}$ because

$f(v^{(1)}{}_{11}) = 3,\quad f(v^{(1)}{}_{21}) = 3,\quad f(v^{(1)}{}_{31}) = 3,\quad f(v^{(1)}{}_{42}) = 3,\quad f(v^{(1)}{}_{52}) = 3.$

For c_2 we choose: $v^{(2)}{}_{12,}v^{(2)}{}_{22}, v^{(2)}{}_{32}, v^{(2)}{}_{42}$ because

$f(v^{(2)}{}_{12}) = 3, \quad f(v^{(2)}{}_{22}) = 3, \quad f(v^{(2)}{}_{32}) = 2, \quad f(v^{(2)}{}_{42}) = 2,$

For c_3 we choose: $v^{(3)}{}_{21}$ because $f(v^{(2)}{}_{21}) = 2.$

Thus, determined opening learning scenario s* is equal:

$s^* = << v^{(1)}{}_{31}, v^{(1)}{}_{52}, v^{(1)}{}_{42}, v^{(1)}{}_{11}, v^{(1)}{}_{21} >, < v^{(2)}{}_{12,}v^{(2)}{}_{22}, v^{(2)}{}_{42}, v^{(2)}{}_{32} >, < v^{(3)}{}_{21} >>$

5. Conclusion and Future Works

Students choose systems for distance education heartily if they are offered a learning path suitable for their preferences, learning styles, abilities etc. Thus, very important task is to determine an opening learning scenario. Proposed procedure allows learning the most adequate material in the best order suitable for an individual.

In this paper knowledge structure and learning scenario are defined. The proposed knowledge structure makes creating different courses and learning scenarios possible and allows adapting suitable version of lesson such as textual, graphical, interactive etc. for student's preferences.

Future tests of the proposed method in comparison with different procedures are planned. The next step will be implementing an e-learning system and research influence of using described algorithm on the efficiency of a learning process. Proposed solution could be helpful for designers of intelligent tutoring systems.

References

[1] P. Brusilovsky, Adaptive and Intelligent Technologies for Web-based Education, In: C. Rollinger and C. Peylo (eds.) *Künstliche Intelligenz*, Special Issue on Intelligent Systems and Teleteaching 4 (1999), 19-25.

[2] P.Brusilovsky, Ch.Peylo, Adaptive and Intelligent Web-based Educational Systems, *International Journal of Artificial Intelligence in Education* 13 (2003), 156–169.

[3] M. Grigoriadou, K. Papanikolaou, H. Kornilakis, G. Magoulas, INSPIRE: An INtelligent System for Personalized Instruction in a Remote Environment, *Lecture Notes In Computer Science*. 2266 (2001), 215-225.

[4] A. Kozierkiewicz, Content and structure of learner profile in an intelligent E-learning system, N.T. Nguyen, G. Kolaczek, B. Gabrys (Eds.): *Knowledge Processing and Reasoning for Information Society*, EXIT Warsaw.

[5] E. Kukla, N.T. Nguyen, C. Daniłowicz, J. Sobecki, M. Lenar, A model conception for optimal scenario determination in an intelligent learning system, *ITSE - International Journal of Interactive Technology and Smart Education* 1 (2004), 171-184.

[6] E. Melis, E. Andres, J. Budenbender, A. Frischauf, G. Goguadze, P. Libbrecht, M. Pollet, C. Ullrich, ActiveMath: A Generic and Adaptive Web-Based Learning Environment, *International Journal of Artificial Intelligence.*

[7] N.T. Nguyen, *Advanced Methods for Inconsistent Knowledge Management*, Springer-Verlag, New York, 2008

[8] G. Weber, P. Brusilovsky, ELM-ART: An Adaptive Versatile System for Web-based Instruction, *International Journal of Artificial Intelligence in Education* 12 (2001),.351-384.

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-214

Acquisition of a New Type of Lexical-Semantic Relation from German Corpora

Lothar LEMNITZER [a] and Piklu GUPTA [a] and Holger WUNSCH [a]
[a] *Seminar für Sprachwissenschaft, Universität Tübingen, Germany*

Abstract. In this paper we will report on work in progress towards increasing the relational density of the German wordnet. It is also an experiment in corpus-based lexical acquisition. The source of the acquisition is a large corpus of German newspaper texts. The target is the German wordnet (GermaNet). We acquire a new type of lexical-semantic relation, i.e. the relation between the verbal head of a predicate and the nominal head of its argument. We investigate how the insertion of instances of this relation into the German wordnet GermaNet affects the neighbourhood of the nodes which are connected by an instance of the new relation. Special attention is given in this paper to the language-specific aspects of the acquisition process.

Keywords. Wordnets, Lexical Acquisition, Semantic Relatedness

Introduction

In this paper we report on a lexical acquisition experiment. The sources of the acquisition are two large, syntactically annotated German corpora: a newspaper corpus and the German Wikipedia. We use these corpora and draw on their syntactic annotation to extract pairs of verbs and their arguments in order to include these relations in the German wordnet.

Wordnets are a valuable lexical-semantic resource used with many NLP applications and techniques. Major use cases for wordnets are applications in which the detection of semantic similarity or semantic relatedness play an important role, e.g. lexical chaining and semantic information retrieval[1].

The main characteristics of wordnets are the organisation of lexical units into synsets and the connection of both lexical units and synsets by lexical-semantic relations. In this paper we will focus on one particular lexical-semantic relation. The types of lexical semantic relations which can be found in almost all wordnets are a) synonymy; b) antonymy; c) hypernymy / hyponymy; d) holonymy / meronymy; e) troponymy ; f) causation; g) entailment; h) pertainymy. We call these relations the classical relation types. These relation types are also used in GermaNet.

[1] Morato et al. [1] give an overview of applications of wordnets in general. Gurevych et al. [3], [2] describe the use of GermaNet in semantic information retrieval. Cramer and Finthammer [4] discuss issues of the use of GermaNet for lexical chaining.

NLP applications such as the one described above draw on the existence of relations between lexical units and synsets (in the following: lexical objects) in order to measure semantic relatedness or similarity. These applications view wordnets as graphs in which the lexical objects are the nodes and the relations between these objects are the edges. Semantic relatedness or similarity between lexical objects is measured by the length of the shortest path between two lexical objects[2].

It has been shown recently that wordnets suffer from the relatively small number of relation instances between their lexical objects (cf. Boyd-Graber et al. [6]). It is assumed, e.g. by Boyd-Graber et al., that applications in NLP and IR, in particular those relying on word sense disambiguation, can be boosted by a lexical-semantic resource with a higher relational density and, consequently, shorter average paths between the lexical objects. This situation also applies to the German wordnet. It consists of 57,700 synsets with 81,800 lexical units. These lexical objects are connected by only 3,678 lexical relations between lexical units and 64,000 conceptual relations between synsets. Even if we count lexical objects which are related indirectly through an intermediate node, the network is not very dense, and paths between words become very long.

We therefore decided to acquire new lexical relations from text corpora in order to make the network denser. In this paper, we report on research in the lexical acquisition of a new type of relation from large annotated German corpora. We focus on the relation between the verbal heads of predicates and the nominal heads of their arguments.

Special attention was given in our investigations to the question of how the insertion of instances of this relation into the German wordnet GermaNet affects the neighbourhood of the nodes which are connected by an instance of the new relation. In particular, we will assess the decrease in the sum total of path lengths which connect the newly related nodes and the nodes which are in the neighbourhood of these nodes. To achieve this, we calculate the shortest paths between any two lexical objects and present the average sum total for each of a set of 100 newly inserted relations. We compare the measures for the original wordnet and the wordnet with the new relation instance added. The impact of the new relation instances on the sum and distribution of path lengths serves as a benchmark for the efficiency of several acquisition methods. We consider this to be a relevant benchmark because the decrease in path lengths positively affects methods of measuring semantic relatedness which are based on these path lengths .

The rest of the paper is organised as follows: we start with an overview of related work; in section 2 we describe the corpora which we have used for our acquisition experiments, in section 3 the acquisition methods are described. Section 4 is devoted to the experiments with the extended GermaNet and their outcomes. We finish the paper with a section in which we draw conclusions and outline future work.

1. Related Work

The work on which we report in our paper is an example of acquisition of lexical-semantic descriptions with the aim of structurally enriching a lexical-semantic resource. Research in the acquisition and integration of new synsets aims to reduce the amount of time-consuming and error-prone manual work required to extend these resources. Snow

[2]Budanitsky and Hirst [5] give an overview of methods by which the shortest path between any two lexical objects is calculated.

et al. [7] and Kim Sang [8] present highly efficient approaches to this task. They exploit the fact that taxonomic relations between lexical objects are reflected by distributional patterns of these lexical objects in corpora. This kind of research, however, is not in the scope of this paper. Instead, we deal with the introduction of relation instances between synsets which are already included in the wordnet, and in particular with instances of a new type of relation. This relation type connects verbal predicates and their nominal arguments.

Research in the (semi-)automatic detection and integration of relations between synsets has boomed in recent years. These activities can be seen as a response to what Boyd-Graber et al. [6] identify as a weakness of the Princeton WordNet: "WordNet, a ubiquitous tool for natural language processing, suffers from sparsity of connections between its component concepts (synsets)." Indeed, the number of relation instances in GermaNet is surprisingly low and needs to be increased.

An entire task of the latest SEMEVAL competition has been dedicated to the detection and classification of semantic relations between nominals in a sentence (cf. Girju et al. [9]). This line of research, however, is targeted at the detection of classical lexical-semantic relations such as hyperonymy.

Some effort has been made to introduce non-classical, cross-category relations into wordnets. Boyd-Graber et al. [6] introduce a type of relation which they call "evocation". This relation expresses that the source concept as a stimulus evokes the target concept. In other words, this is a mental relation which cuts across all parts of speech. This makes the approach different from ours, since we use corpus data instead of experimental data and we acquire what is in the texts rather than what is in the human mind. The relation we introduce is syntactically motivated, which is not the case in the experiment on which Boyd-Graber et al. report.

Amaro et al. [10] intend to enrich wordnets with abstract predicate-argument structures, where the arguments are not real lexical units or synsets but rather abstract labels like INSTRUMENT. They aim at a lexical-semantic resource which supports the semantic component of a deep parser. Therefore they introduce a level of abstraction in the categorisation of the arguments. This is not what we intend to do.

Yamamoto and Isahara [11] extract non-taxonomic, in particular thematic relations between predicates and their arguments. They extract these related pairs from corpora by using syntactic relations as clues. In this respect their work is comparable to ours. Also their aim, i.e. improving the performance of information retrieval systems with this kind of relation, is comparable to ours. However, they do not use the extracted word sets to include them in a wordnet.

Closest to ours is the work of Bentivogli and Pianta [12]. Their research is embedded in the context of machine translation. Seen from this perspective, the almost exclusive representation of single lexical units and their semantic properties is not satisfying. They therefore propose to model the combinatoric idiosyncrasies of lexical units by two new means: a) the phraset as a type of synset which contains multi-word lexical units and b) syntagmatic relations between verbs and their arguments as an extension of the traditional paradigmatic relations. Their work, however, focuses on the identification and integration of phrasets. They only resort to syntagmatic relations where the introduction of a phraset would not otherwise be justified. We take the opposite approach in that we focus on the introduction of instances of the verb-argument relation and resort to the intro-

duction of phrases only in those cases where it is not possible to ascribe an independent meaning to one of the lexical units (cf. section 3).

2. The Corpora

For the acquisition experiments we use two large German corpora: a) the *Tübingen Partially Parsed Corpus of Written German* (TüPP-D/Z) and b) The German Wikipedia. The first corpus contains approximately 11.5 million sentences and 204 million lexical tokens, the second corpus contains 730 million lexical tokens. The corpora were respectively automatically annotated using the cascaded finite state parser *KaRoPars* which has been developed at the University of Tübingen (cf. Müller [13]) and a modified version of the BitPar PCFG parser, cf. Schmid [14] and Versley [15].

The results of the automatic linguistic analysis, however, have not been corrected manually, due to the size of the corpora. Therefore we have to choose an acquisition method which is not sensitive to errors in the annotation.

The sentence presented in **figure 1** translates to "We need to sell the villas in order to pay the young scientists" where the accusative object *Villenverkauf* means "sale of the villas". The sentence in **figure 2** translates to "He gets most inspiration from his father" where the accusative object *Inspiration* occurs before the subject *er* ("he").

The parsers analyse and mark four levels of constituency:

1. The lexical level. For each word in both examples, the part of speech is specified using the Stuttgart-Tübingen-Tagset (STTS, cf. [16]) which is a de-facto standard for German. The elements of the sentence in **figure 2** are additionally annotated with some morphological features. In the example, some words are marked as heads of their respective chunks ("HD"), e.g. *Villenverkauf* and *Inspiration*.
2. The chunk level. Chunks are non-recursive constituents and are therefore simpler than phrases. The use of chunks makes the overall syntactic structure flatter in comparison to deep parsing. In **figure 2** we have two noun chunks (*NCX*) and one verb chunk (*VXVF*). These are the categories which we need for our extraction experiments. Of utmost importance is the functional specification of the noun chunks. They are, with regard to the predicate of the sentence, marked as subject noun phrase (in the nominative case, *ON*) and direct object (in the accusative case, *OA*).
3. The level of topological fields. The German clause can structurally be divided into the verbal bracket, i.e. part of the verb in second position and the other part at the end of the clause, while the arguments and adjuncts are distributed more or less freely over the three fields into which the verbal bracket divides the clause: *Vorfeld* (label: VF) (in front of the left verb bracket), *Mittelfeld* between the two brackets and *Nachfeld* (NF) following the right bracket.
4. The clausal level. The clause in **figure 2** is labelled as a simplex clause (*SIMPX*). In the example in **figure 1** we have an "S" as the root of the sentence.

The parse tree which is generated by KaRoPars (i.e. the example in **figure 1**) is quite flat. Due to limitations of the finite state parsing model, the syntactic relations between the chunks remain underspecified. Major constituents however are annotated with grammatical functions in both cases. This information, together with the chunk and head information is sufficient to extract the word pairs which we need for our experiments.

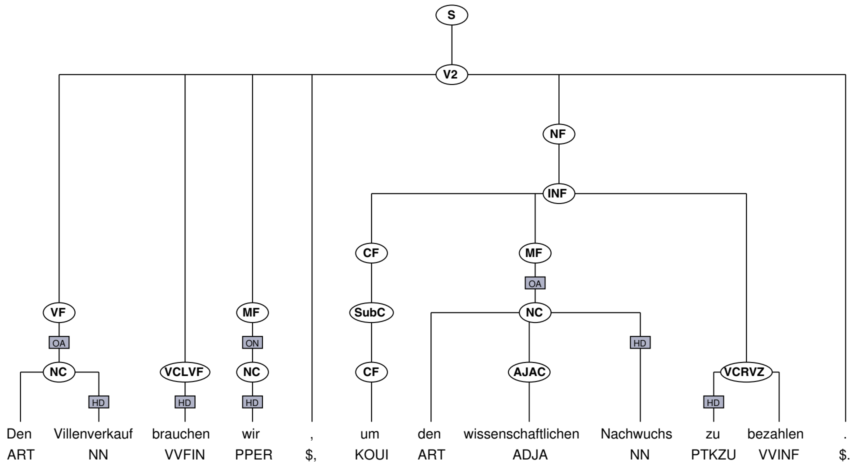


Figure 1. Parse tree of a TüPP-D/Z sentence

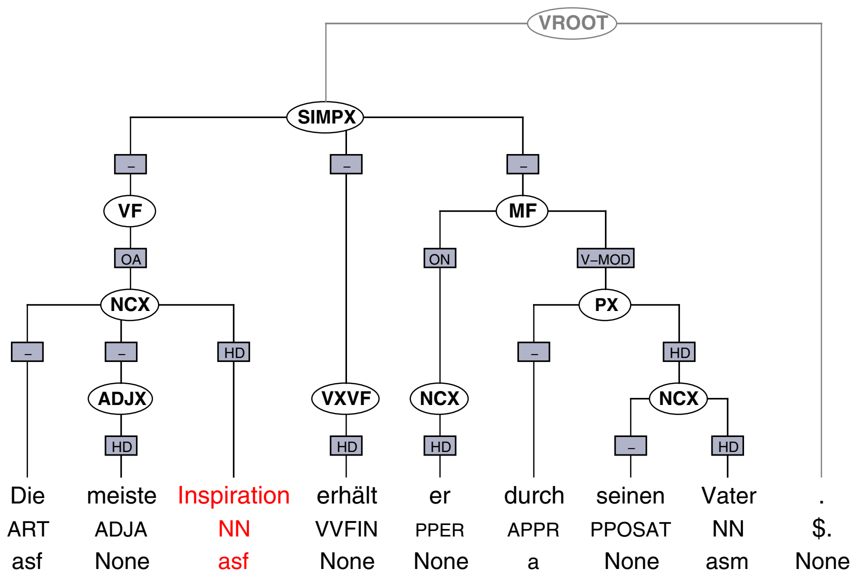


Figure 2. Parse tree of a Wikipedia sentence

3. Acquisition Methods Used

Starting from the linguistically analysed and annotated corpora which we have described above, we extracted two types of syntactically related word pairs: a) verb-subject (e.g. *untersuchen, Arzt – examine, doctor*) and b) verb-direct object (e.g. *finden, Weg,– find, way*)[3]. While the spectrum of possible subjects of a verb turned out to be much broader

[3]The acquisition methods above do not necessarily hinge on the existence of a (partially) parsed corpus. As an approximation, if no such corpus or tools for linguistic annotation are available, pairs of adjacent words can be extracted, after having removed stopwords from the text. This, however, leads to more noise in the data and therefore to a more time-consuming cleaning of the data.

and heterogeneous, verb-object pairs were more readily identifiable and recurrent. Even in scenarios in which associations are arrived at on the basis of evocation, it is interesting to observe that, for instance, Schulte im Walde [17] found a higher number of associations arrived at by humans between verbs and their direct objects than between verbs and their transitive or intransitive subjects. Therefore we focused our work on the analysis of verb-object pairs. By these word-pairs, we capture a subset of the predicate-argument relations in a text, namely those cases in which the predicate is a simple finite verb and the argument is a simple NP in direct object position. More complex structures are beyond the scope of this approach.

In order to rank the word pairs, we measured their collocational strength, which we consider to be a good indicator for their semantic relatedness.

Two common measures – mutual information (MI), cf Church et al. [18] and log-likelihood ratio, cf. Dunning [19]) – are used and compared in our experiments. Mutual information can be regarded as a measurement of how strongly the occurrence of one word determines the occurrence of another; it compares the probability of, for example, two words occurring together with the probability of observing them independently of one another. Log-likelihood ratio compares expected and observed frequencies as might be expressed in a contingency table, a 2 by 2 table where the four cell values represent frequency of word x occurring with word y, x and not y, not x and y and finally not x and not y, i.e. the number of observations where neither word appears. Mutual information seems to be the better choice for the extraction of complex terminological units, due to the fact that it assigns higher weights to word pairs the parts of which occur rather infrequently throughout the corpus. Log likelihood ratio, in contrast, is not sensitive to low occurrence rates of the individual word and is therefore more appropriate for finding recurrent word combinations. This coincides with the findings which are reported by Kilgarriff [20] and, for German, by Lemnitzer and Kunze [21]. Nevertheless, we want to test this assumption independently by measuring the average gain in path lengths for the highest ranked word pairs according to either method.

In order to compare both measures, we took the first 100 entries from the lists of MI-ranked and G^2-ranked word pairs and inserted them one by one into GermaNet. Before selecting these word pairs, we had to clean the lists of pairs which we did not want to insert into the wordnet. There are some word pairs which, for different reasons, we do not want to insert into GermaNet:

1. pairs with wrongly assigned words due to errors in the linguistic annotation. As has been stated above, the linguistic annotation was fully automatic and therefore produced errors. Some of these errors are recurrent and lead to word pairs which are highly ranked. For example, in the sentence *er will Bananen kaufen* (he wants to buy bananas) the modal verb is marked as the verbal head. This word is extracted together with the correctly marked nominal head *Bananen* and forms a pair of base forms *wollen, Bananen*. This is not a relation which we want to insert into GermaNet.
2. pairs with words which did not have an entry in GermaNet. It was not our primary goal to extend the lexical base of GermaNet. We therefore ignored word pairs where at least one element could not be found in the wordnet.

3. word pairs which are fixed expressions or parts of them. It is well-known that many idioms expose syntactic as well as semantic irregularities. In particular, it is impossible to assign a standard meaning to the individual words. The idiom *den Löffel abgeben* (to hand in the spoon), for example, has nothing to do with cutlery, but, as a whole, is a colloquial expression for *sterben* (to die).
4. support verb constructions were also discarded. It has been argued convincingly by Storrer [22], that some types of support verb constructions and their verbal counterparts, e.g. *eine Absage erteilen* and *absagen* (lit. to give a rejection, to reject), show subtle differences in their use and therefore support verb constructions merit an independent status as complex lexical units. Besides that, it is often very difficult to assign a meaning to the verbal part of the construction, which is due to the mere supporting function of this part.

For these reasons, we consider it to be inappropriate to represent (semi-)fixed expressions by relating their elements. We still have to find a way of encoding these complex lexical elements in GermaNet, but this is outside the scope of this paper.

The 100 highest ranked of the remaining word pairs were inserted into the wordnet manually. Inserting the word pair involved a manual disambiguation step: all words were mapped to the correct synsets. Semi-automatic insertion of the new relation instances would require reliable word sense disambiguation which is not yet available for German. In the following we report on experiments in which we calculated the local impact of the new relation instances[4].

4. Experiments

4.1. Local effects of adding relations

In our experiments, we wanted to measure the impact of newly introduced relations between verbs and nouns on the path lengths between the related words and the words which are in their neighbourhood.

To achieve this, we first had to make a structural change to GermaNet. In the current version of the wordnet, the nominal and the verbal part are not connected at all. There is no "supernode" which overarches the so-called unique beginners, top nodes that are not connected to each other which correspond to the most general concepts in their respective lexical fields. We introduced an artificial node that connects all subgraphs of GermaNet. In this new version, there is at least one path from every synset to any other synset.

Next, we introduced the relation instances which connect the lexical objects which we acquired by the collocational analyses. We inserted these relations one by one. The settings for measuring the impact of each new relation are as follows: let s_1 and s_2 be two synsets and $R(s_1, s_2)$ the new relation instance connecting the two synsets. Further, let SP_b be the shortest path between s_1 and s_2 before the insertion of $R(s_1, s_2)$ and let SP_a be the shortest path between s_1 and s_2 after the insertion of $R(s_1, s_2)$. By definition, the length of the shortest path between s_1 and s_2 after the insertion of (SP_a) is 1 (see

[4]We also measured the global impact of the new relations, i.e. the impact of these links on the overall reduction of paths length between any two nodes. There are, however, no visible effects and the selected word pairs do not have any impact which is different from the impact of the same number of relations inserted between randomly chosen lexical object. For details, cf. Lemnitzer et al. [23]

Table 1. Cumulative path length reduction, average of 100 word pairs for both MI a nd G^2.

method	average PR value
MI	2762.04
G^2	15867.38

figure 3). We calculate the path reduction PR_{s_1,s_2} as the result of $SP_b - SP_a$. We now take S_1 and S_2, the sets of all synsets which are in the two subtrees rooted by s_1 and s_2 respectively; in other words, we take all the hyponyms, the hyponyms of these hyponyms and so forth. We calculate the path reduction PR_{s_m,s_n} for each pair $s_m \in S_1, s_n \in S_2$. The sum of all path reduction values is the local impact caused by the new relation instance. We calculated the sum total of the path reduction values for the 100 most highly ranked pairs according to the mutual information and the log-likelihood statistics. **Table 1** shows the average cumulative path reduction value for both statistics.

From these figures we can infer that: a) there is a considerable local impact of the new relation instances, which is what we wanted to achieve; b) the impact of the word pairs extracted by log-likelihood ratio is much higher than that of the pairs extracted by mutual information. This confirms our assumption about the superiority of log-likelihood ratio for our acquisition task.

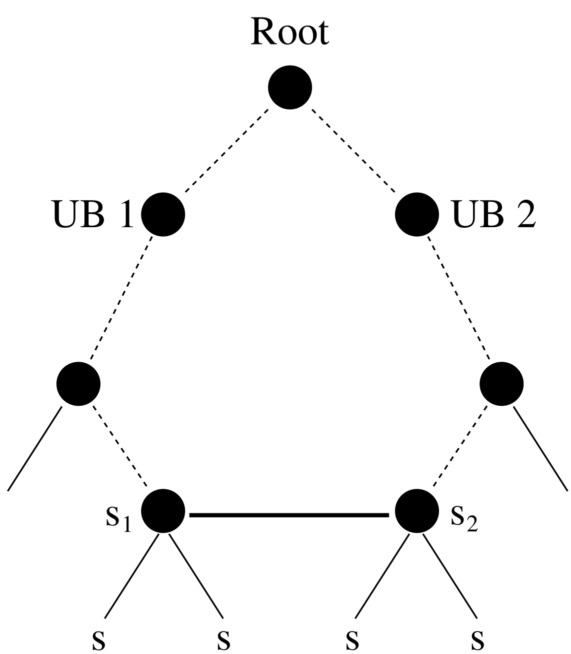


Figure 3. Local path reduction between two synsets s_1 and s_2. The dashed path is the old path, the new relation $R(s_1, s_2)$ is depicted by the thick line between s_1 and s_2.

5. Conclusions and Future Work

We have shown that the insertion of new, cross-categorial relation instances has a measurable local impact. Words which have been far away from each other, but are found to co-occur in our corpora, now become neighbours. We expect that the linking of words of different categories will improve the usefulness of wordnets for applications such as information retrieval and lexical chaining, among others. We also expect an impact on the task of word sense disambiguation. Words might better be disambiguated by *the company they keep*.

Most of the semantic relatedness measures reported by Budanitsky and Hirst [5] are, however, not sensitive to cross-categorial relations. We therefore see more potential in a combination of an extended wordnet with measuring semantic relatedness by random graph walks (cf. Hughes and Ramage [24] for details).

In our experiments, we base our new relations on lexical objects which we found in our corpora. We do not make any effort to abstract away from the individual words and to encode these relations as selectional restrictions or selection preferences of verbs, cf. Agirre and Martinez [25] for such an approach. We consider the abstraction of selectional preferences from corpus-attested word pairs as the second step. This step can be performed on the data which we insert into our wordnet. For example, for any number of nouns $N_1 \dots N_n$ which are related to a particular verb, or for any subset of these, one can look for the lowest common subsumer and check whether this lexical object is an appropriate abstraction from the individual nouns. But we also foresee a number of applications where such an abstraction is not needed.

We have currently inserted around 10 000 new relation instances of the verb-object type. In the near future we will investigate the impact of the new relation on the performance of semantic information retrieval and on anaphora and coreference resolution. If this approach proves to be useful for applications using GermaNet, and we expect it to be, many relation instances can be transferred to other wordnets via the Interlingual Index which connects these wordnets.

Acknowledgements

The research which is described in this paper has been funded by a grant from the "Deutsche Forschungsgemeinschaft" for the project "Semantic Information Retrieval". We are grateful for the support.

References

[1] J. Morato, M. Marzal, J. Lloréns, and J. Moreiro. WordNet Applications. In P. Sojka, K. Pala, P. Smrž, C. Fellbaum, and P. Vossen, editors, *Proceedings of the Second International WordNet Conference—GWC 2004*, pages 270–278, Brno, Czech Republic, 2003. Masaryk University Brno, Czech Republic.

[2] I. Gurevych. Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP'2005)*, Jeju Island, Republic of Korea, Oct. 2005.

[3] I. Gurevych, C. Müller, and T. Zesch. What to be? - electronic career guidance based on semantic relatedness. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 1032–1039, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[4] I. Cramer and M. Finthammer. An Evaluation Procedure for Word Net Based Lexical Chaining: Methods and Issues. In *Proc. Global WordNet Conference 2008*, pages 120–146, Szeged, 2008.

[5] A. Budanitsky and G. Hirst. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

[6] J. Boyd-Graber, C. Fellbaum, D. Osherson, and R. Schapire. Adding Dense, Weighted Connections to WordNet. In *Proceedings of the Third International WordNet Conference*, Masaryk University, Brno, Czech Republic, 2006.

[7] R. Snow, D. Jurafsky, and A. Y. Ng. Semantic taxonomy induction from heterogenous evidence. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 801–808, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[8] E. Tjong Kim Sang. Extracting hypernym pairs from the web. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 165–168, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[9] R. Girju, P. Nakov, V. Nastase, S. Szpakowicz, P. Turney, and D. Yuret. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 13–18, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[10] R. Amaro, R. P. Chaves, P. Marrafa, and S. Mendes. Enriching wordnets with new relations and with event and argument structures. In *Proc. CICLing 2006*, pages 28–40, 2006.

[11] E. Yamamoto and H. Isahara. Extracting word sets with non-taxonomical relation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 141–144, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[12] L. Bentivogli and E. Pianta. Extending WordNet with Syntagmatic Information. In P. Sojka, K. Pala, P. Smrž, C. Fellbaum, and P. Vossen, editors, *Proceedings of the Second International WordNet Conference—GWC 2004*, pages 47–53, Brno, Czech Republic, 2003. Masaryk University Brno, Czech Republic.

[13] F. H. Müller. Stylebook for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z), 2004.

[14] H. Schmid. Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, 2004.

[15] Y. Versley. Parser Evaluation across Text Types. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, Barcelona, Spain, 2005.

[16] A. Schiller, S. Teufel, C. Thielen, and C. Stöckert. Guidelines für das Taggen deutscher Textcorpora mit STTS. 1999.

[17] S. Schulte im Walde. Can Human Verb Associations help identify Salient Features for Semantic Verb Classification? In *Proceedings of the 10th Conference on Computational Natural Language Learning*, New York City, NY, 2006.

[18] K. Church, W. Gale, P. Hanks, and D. Hindle. Using Statistics in Lexical Analysis. In U. Zernik, editor, *Lexical acquisition: exploiting on-line resources to build a lexicon*, pages 115–164. Laurence Erlbaum Associates, Hillsdale, NJ, 1991.

[19] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[20] A. Kilgarriff. Which words are particularly characteristic of a text? A survey of statistical approaches. In *Proceedings of AISB Workshop on Language Engineering for Document Analysis and Recognition*, pages 33–40, Falmer, Sussex, 1996.

[21] L. Lemnitzer and C. Kunze. *Computerlexikographie*. Gunter Narr Verlag, Tübingen, 2007.

[22] A. Storrer. Funktionen von Nominalisierungsverbgefügen im Text. Eine korpusbasierte Fallstudie. In K. Proost and E. Winkler, editors, *Von Intentionalität zur Bedeutung konventionalisierter Zeichen. Festschrift für Gisela Harras zum 65. Geburtstag*, pages 147–178. Gunter Narr, Tübingen, 2006.

[23] L. Lemnitzer, H. Wunsch, and P. Gupta. Enriching GermaNet with verb-noun relations – a case study of lexical acquisition. In *Proc. LREC 2008*, Marrakech, 2008.

[24] T. Hughes and D. Ramage. Lexical Semantic Relatedness with Random Graph Walks. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Compu-*

tational Natural Language Learning, pages 581–589, Prague, Czech Republic, 2007. Association for Computational Linguistics.

[25] E. Agirre and D. Martinez. Learning class-to-class selectional preferences. Proceedings of the Workshop 'Computational Natural Language Learning' (CoNLL-2001). In conjunction with ACL 2001/EACL 2001.

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-225

Problems of Data Quality in an Integrated Cadastral Information System

Dariusz CIEŚLA[a1], Zbigniew TELEC[a], Bogdan TRAWIŃSKI[b], and Robert WIDZ[a]
[a]*Intergraph Polska Sp. z o.o., Poland*
[b]*Wrocław University of Technology, Institute of Applied Informatics, Poland*
e-mail: dariusz.ciesla@intergraph.com, zbigniew.telec@intergraph.com, bogdan.trawinski@pwr.wroc.pl, robert.widz@intergraph.com

Abstract. Cadastral systems belong to the most important public systems, since they provide necessary information for economic planning, spatial planning, and tax calculation, real estate denotation in perpetual books, public statistics, and real estate management as well as farm registration. They also provide source data to systems such as IACS and its main component LPIS. Quality of cadastral data has a major impact on the functioning many other systems. Legacy cadastral systems were developed using different data structures and various computer technologies. The quality of those legacy systems was sufficient for accomplishing everyday tasks but it was too low in the case of transferring data into other information systems. A practical approach to assure data quality in the process of transferring descriptive and geometric data from legacy systems into the Kataster OnLine - a modern integrated cadastral information system is presented in the paper.

Keywords. cadastral system, data quality, Kataster OnLine

Introduction

Cadastre systems are mission critical systems designed for the registration of parcels, buildings and apartments as well as their owners and users and they are comprised by the governmental information resources. Those systems have complex data structures and sophisticated procedures of data processing. Spatial data are of strategic importance for any country, because land information is involved in 80% of administration activity, and each government is responsible for the provision of its infrastructure. In addition, it is a necessity to secure land property rights and the certainty of title, the efficient management of state land assets and the maintenance of publicly available geodetic and cartographic information. Cadastre systems belong to the most important public systems, since they provide necessary information for economic planning, spatial planning, and tax calculation, real estate denotation in perpetual books, public statistics, and real estate management as well as farm registration [7], [8]. They also provide source data to systems such as IACS (Integrated Administration and Control System) and its main component LPIS (Land Parcel Identification System) – the systems of EU payments to agricultural land. The list of major users of spatial information system is given in Figure 1.

[1] Corresponding author: Dariusz Cieśla, Intergraph Polska Sp. z.o.o., ul. Domaniewska 52, 02-672 Warszawa, Poland; E-mail: dariusz.ciesla@intergraph.com.

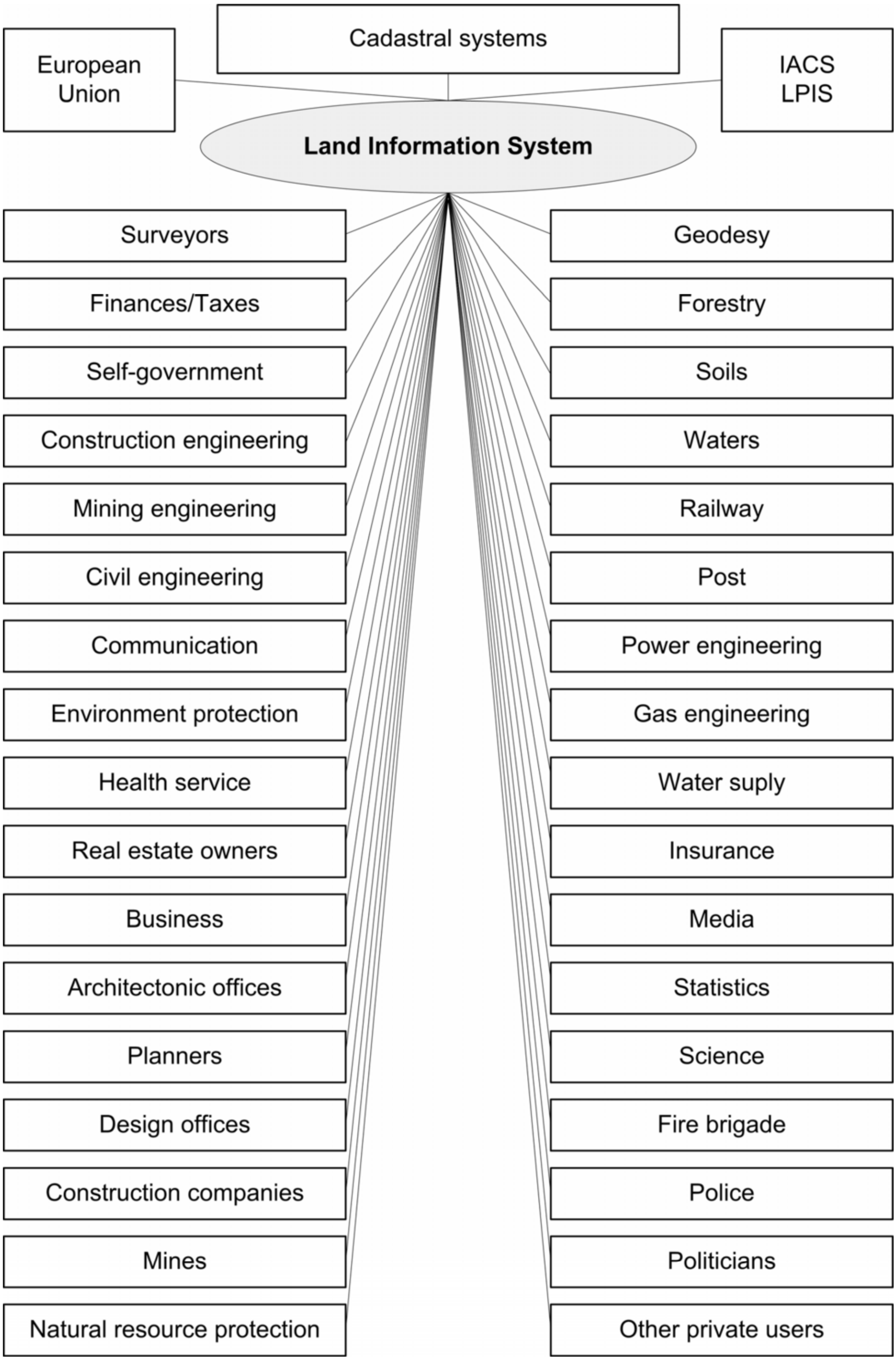


Figure 1. Users of Land Information System

The problem of data quality is crucial for cadastral information systems. Data quality is related to the satisfaction of the intended use and data must be accurate, timely, relevant, and complete [4]. At present, the definitions of good quality data focus mainly on its consumers and its use [9], [10], and they often take the form of the question to what extent data satisfy the requirements of their intended use. The primary job of the data quality investigation is to identify specific instances of wrong values [1]. Examples of metrics of data quality that can be gathered are: key violations (duplicates, orphans), rows with at least one wrong value, percent of records never returned as answers, object fan-in, and fan-out, number of used attributes per object. Many companies developed to assess data quality the following dimensions [9]: accessibility, believability, completeness, integrity, flexibility, free-of-error, timeliness, simplicity, and comprehensibility. By these dimensions, three functional forms help in practice: simple ratio, min or max operators, and weighted average [5], [11].

Most often the process of data quality monitoring is the introductory part of a more general process of data quality improvement [2], [11]. It is focused on the analysis of defect occurrence in order to remove them automatically. Human verification and approval is needed to solve many problems [3], thus it is suggested to distinguish clearly both processes.

Quality of cadastral data has a major impact on functioning many other systems. Legacy cadastral systems were developed using different data structures and various computer technologies. The quality of those legacy systems was sufficient for accomplishing everyday tasks but it was too low in the case of transferring data into other information systems. Data exchange was also hindered because of different standards of data gathering in cadastral systems. New regulations introduced in Poland in 2003 aimed at unifying the standard of data gathered in cadastral systems, improving the quality of descriptive and geometric data, removing the discrepancy between descriptive and geometric data by establishing the unified object model of cadastral data, introducing unified standard of cadastral data exchange called SWDE.

First improvement of data quality was achieved during the process of exporting data from cadastral systems into the IACS - Integrated Administration and Control System that is centralized at the level of the country. Many errors were removed and lacks were supplemented in cadastral data, and their structure was adjusted to the SWDE exchange standard. The main difficulties lied in transferring legacy data structures into object ones and in ensuring permanent object identifiers. In fact only totally new systems developed from scratch are able to meet all contemporary data quality requirements. The Kataster OnLine is the first such system.

A practical approach to ensure data quality in the process of transferring descriptive and geometric data from a legacy system into the Kataster OnLine - a modern integrated cadastral information system is presented in the paper.

1. Kataster OnLine - Integrated Cadastral Information System

Kataster OnLine system is a modern cadastral information system, which integrates descriptive and geometric data in one database. It is a new generation system, based on new trends and technologies in constructing information systems, such as Oracle database system and GIS tools like GeoMedia WebMap. It has been constructed using web technology making it possible to access and update cadastral data remotely using thin client application. The advantage of such solution is that data can be modified

online, and the database at any moment is up-to-date and consistent. The genesis and architecture of the first version of the Kataster OnLine system was described in [6].

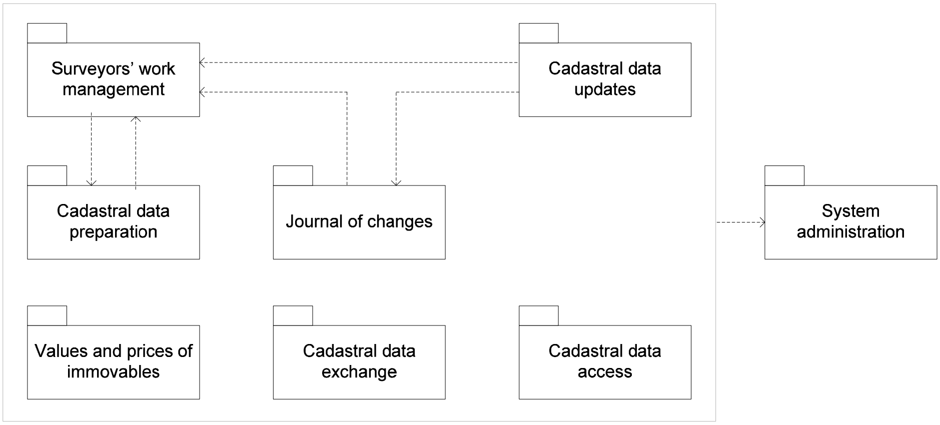


Figure 2. Module structure of the Kataster Online system

As the result of thorough analysis a series of processes accomplished in information centres maintaining cadastral registers have been isolated and described. These processes have been reflected in the Kataster OnLine system in the form of scenarios. The scenarios guide users performing different tasks, allowing them to input strictly determined data into the system form by form. Therefore, the possibility of making errors has been limited. All scenarios have been grouped into 8 modules, which provide the functionality of the system. The module structure of the Kataster OnLine system is shown in Figure 2 and it comprises following modules:

- ***Surveyors' work management (SWM)*** - comprises functions for registering and updating geodetic works, reservation of the numbers of parcels, buildings and boundary corner points.
- ***Cadastral data preparation (CDP)*** – is designed for updating geometric data through parcel division, amalgamation and demarcation, inserting, deleting and modifying data of buildings and other registration objects. Available are also tools for preparing data for surveyors, for importing points and pickets from files containing coordinates, accomplishing geodetic calculations, for controlling data correctness.
- ***Journal of changes (JC)*** – is an office module devoted to register documents constituting a legal basis of changes and to determine the scope of changes.
- ***Cadastral data updates (CDU)*** – allows to modify data of cadastral objects and subjects using especially designed mechanism of long transactions. At the end of each transaction, the change report in the form of a list of updated objects is displayed and then whole change can be approved by a user and committed.
- ***Cadastral data access (CDA)*** – enables registered users to search and view cadastral data in the form of reports, indexes, lists and various extracts and maps.

- ***Cadastral data exchange (CDE)*** – is designed for export of data concerning selected area to the standard SWDE exchange format, complete as well as differential one.
- ***Values and prices of immovables (VPI)*** – is designed for updating geometric data through parcel division, amalgamation and demarcation, inserting, deleting and modifying data of buildings and other registration objects. Available are also tools for importing points and pickets from files containing coordinates
- ***System administration (SA)*** – enables system administrators to set system parameters, to maintain user accounts and to grant users their rights to access individual modules, scenarios, legends, work areas. Moreover it allows to monitor events occurring in the system as well as trace users' activity.

The Kataster OnLine system is being continuously developed, and next modules of basic map and real estate management are being prepared to incorporate into the integrated system.

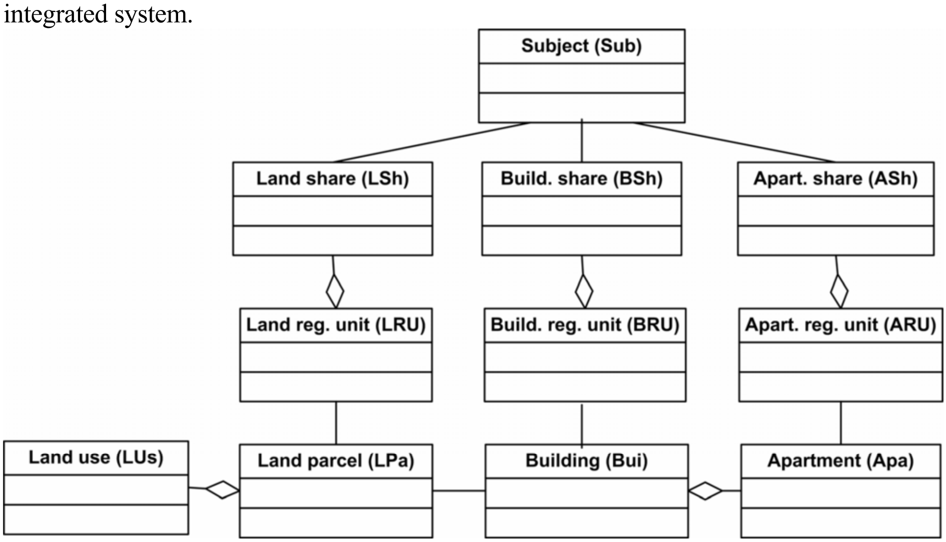


Figure 3. Main classes of cadastral object model

The structure of main cadastral objects of descriptive data applied in the Kataster OnLine system, which conforms with new regulations, is presented in the form of a UML class diagram in Figure 3. The structure of geometric data was also designed in accordance with the regulations. The diagram shows that registration units, containing subjects with their shares as well as land parcels, buildings and apartments respectively, are the main processing units in the system. Following denotations of main data objects are used later in the paper:

- LPa – land parcels,
- Bui – buildings,
- Apa – apartments which comprise residential and non-residential premises,
- LRU – land registration units,
- BRU – buildings registration units,
- ARU – apartment registration units,
- LSh – land shares,

- BSh – building shares,
- ASh – apartment shares,
- Sub – cadastral subjects which comprise physical persons, institutions, marriages and collective subjects,
- LUs – land uses.

2. Data Quality Assurance During the Exchange of Cadastral Systems

The Kataster OnLine is a relatively young system but it can be expected that it will be replacing legacy cadastral systems in Poland in the near feature. The process of data conversion from legacy cadastral systems to the Kataster OnLine system is shown in Figure 4. You can see that descriptive and geometric data are transferred from the source system separately and only in the target system they are integrated into one unified database.

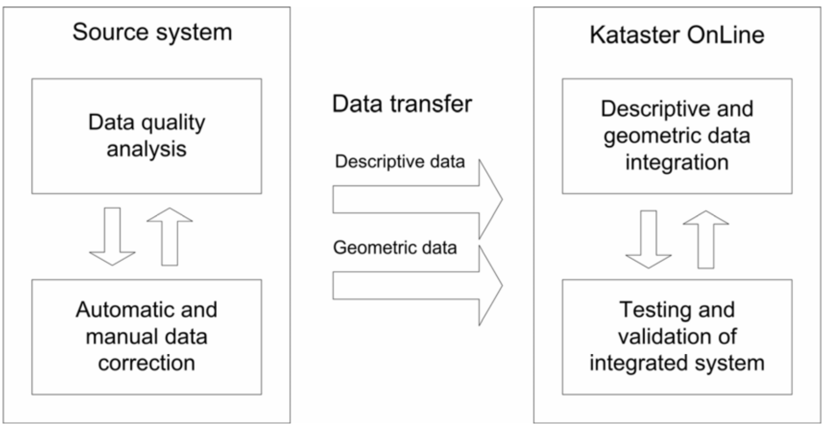


Figure 4. Process of data conversion to the Kataster Online system

The process encompasses the quality assurance procedures both on the side of the source system and the side of the target Kataster OnLine system. It is composed of following steps:

- agreement on conversion procedure and handover of a database,
- data import into a transitory database,
- thorough data quality analysis,
- preparing a report on data quality with the recommendation of remedial measures,
- decision about repair methods and the schedule,
- automatic removal of errors,
- manual correction where administrative procedure is needed,
- data quality re-analysis after the repair,
- data import into a production database,
- integration of descriptive and geometric data,
- database testing and preparing a report on data quality,
- decision on starting exploitation of the system.

In order to illustrate the problems of data quality assurance the case study of the conversion of descriptive and geometric data from legacy cadastral systems to the Kataster OnLine in five information centres in Poland is described below. In Table 1 the numbers of cadastral objects transferred to the Kataster OnLine system in respective centres are presented.

Table 1. Numbers of cadastral objects transferred to the Kataster OnLine system

Object	Centre 1	Centre 2	Centre 3	Centre 4	Centre 5
LPa	57 417	81 302	85 951	239 480	162 538
Bui	20 054	7 983	10 028	239 480	20 210
Apa	11 376	3 247	5 893	10 638	303
LRU	41 337	23 596	58 500	115 404	40 485
BRU	12 991	11 781	24 038	40 574	26 555
ARU	5 063	974	2 354	5 085	71
LSh	11 113	2 888	5 841	7 942	166
BSh	89 596	58 127	111 580	196 615	75 182
ASh	19 458	3 468	12 552	0	229
Sub	17 632	4 970	9 400	14 920	282

During the data quality analysis above 350 elementary tests were performed. Some tests were designed to detect defective or lacking values in determined fields of records containing data of main cadastral objects whereas the other to examine incorrect or lacking references between records of different objects. In the case of geometric data syntactic, semantic, and topological verification of data was accomplished. Figures 5-12 refer to errors found in descriptive data whereas Figures 13-23 concern geometric ones.

In Figure 5 the percentage of main cadastral objects with errors found in descriptive data is presented.

In Figure 6 the distribution of errors in land parcels is shown. Following denotation of errors are used: LPa_1 – erroneous denotation of land use; LPa_2 – archival parcels with erroneous date of archiving; LPa_3 – archival parcels with erroneous parcel number; LPa_4 – current parcels with erroneous parcel number; LPa_5 – current parcels in archival land registration units.

In Figure 7 the distribution of errors in buildings is shown. Following denotation of errors is used: Bui_1 – wall material undetermined; Bui_2 – usable function undetermined; Bui_3 – erroneous building identifier; Bui_4 – erroneous year of construction; Bui_5 – lacking or erroneous date of archiving; Bui_6 – referential errors.

In Figure 8 the distribution of errors in apartments is shown. Following denotation of errors is used: Apa_1 – usable function undetermined; Apa_2 – erroneous or duplicated apartment identifier; Apa_3 – referential errors.

In Figure 9 the distribution of errors in land registration units is shown. Following denotation of errors is used: LRU_1 – LRU without any land share; LRU_2 – LRU without any parcel; LRU_3 – current LRU with only archive land shares; LRU_4 – current LRU with only archive parcels.

In Figure 10 the distribution of errors in building registration units is shown. Following denotation of errors is used: BRU_1 – BRU without any building share; BRU_2 – BRU without any building; BRU_3 – current BRU with only archive building shares; BRU_4 – current BRU with only archive buildings; BRU_5 – BRU left in update mode.

In Figure 11 the distribution of errors in apartment registration units is shown. Following denotation of errors is used: ARU_1 – ARU without any apartment share; ARU_2 – ARU without any apartment; ARU_3 – current ARU with only archive apartment shares; ARU_4 – current ARU with only archive apartments; ARU_5 – ARU left in update mode; ARU_6 – lacking identifier.

In Figure 12 the distribution of errors in land shares is shown. Following denotation of errors is used: LSh_1 – erroneous denotation of marriages and collective subjects; LSh_2 – lacking value of share; LSh_3 – lacking or erroneous number of register group; LSh_4 – lacking or erroneous date of archiving; LSh_5 – referential error.

In Figure 13 the distribution of errors in building shares is shown. Following denotation of errors is used: BSh_1 – erroneous denotation of marriages and collective subjects; BSh_2 – lacking value of share; BSh_3 – lacking or erroneous number of register group; BSh_4 – lacking or erroneous date of archiving; BSh_5 – referential error.

In Figure 14 the distribution of errors in apartment shares is shown. Following denotation of errors is used: ASh_1 – erroneous denotation of marriages and collective subjects; ASh_2 – lacking or erroneous number of register group; ASh_3 – referential error.

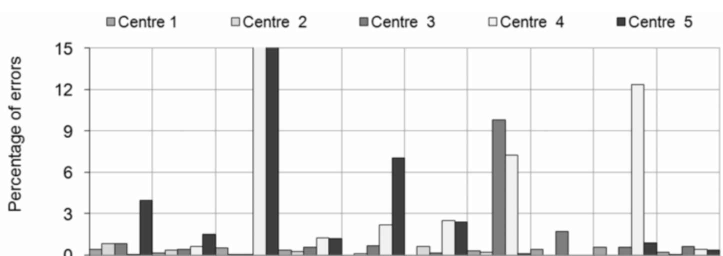
In Figure 15 the distribution of errors in cadastral subjects is shown. Following denotation of errors is used: Sub_1 – lacking surname of a physical person; Sub_2 – lacking or erroneous sex; Sub_3 – lacking type of subject; Sub_4 – lacking or erroneous subject status.

In Figure 16 the percentage of cadastral objects with errors found in geometric data is presented.

In Figure 17 the distribution of errors in land parcels is shown. Following denotation of errors is used: LPa_10 – erroneous drawing of contour; LPa_11 – erroneous drawing of enclave; LPa_12 – lacking coordinates.

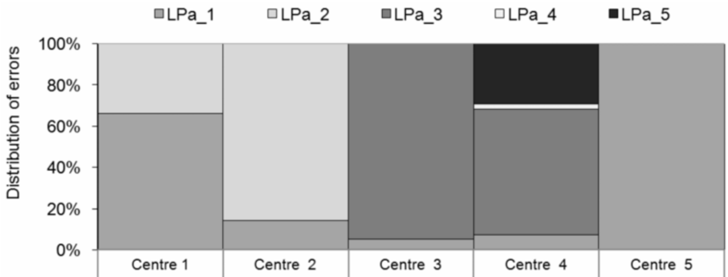
In Figure 18 the distribution of errors in buildings is shown. Following denotation of errors is used: Bui_10 – null area; Bui_11 – erroneous drawing of contour; Bui_12 – erroneous drawing of enclave.

In Figure 19 the distribution of errors in land uses is shown. Following denotation of errors is used: LUs_10 – erroneous drawing of contour; LUs_11 – erroneous drawing of enclave.



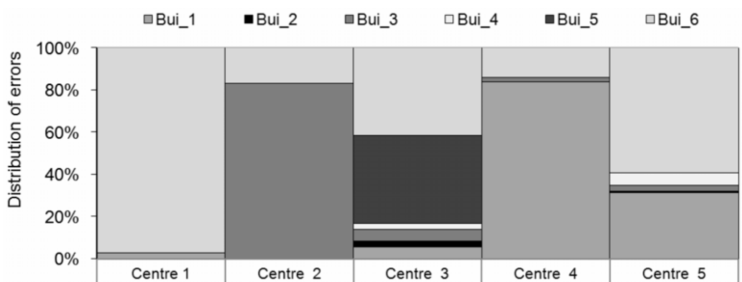
	LPa	Bui	Apa	LRU	BRU	ARU	LSh	BSh	ASh	Sub
Centre 1	0.43	0.18	0.55	0.38	0.02	0.02	0.32	0.45	0.60	0.21
Centre 2	0.83	0.32	0.03	0.23	0.10	0.62	0.20			0.03
Centre 3	0.83	0.38	0.03	0.55	0.68	0.15	9.80	1.68	0.53	0.60
Centre 4	0.06	0.61	20.04	1.24	2.20	2.48	7.27		12.37	0.44
Centre 5	3.98	1.49	32.30	1.21	7.04	2.41	0.11		0.88	0.37

Figure 5. Percentage of objects with errors in descriptive data



	Centre 1	Centre 2	Centre 3	Centre 4	Centre 5
LPa_5				29.3%	
LPa_4				2.4%	
LPa_3			94.2%	60.9%	
LPa_2	33.7%	85.3%			
LPa_1	66.4%	14.3%	5.4%	7.3%	100.0%

Figure 6. Distribution of errors in land parcels (descriptive data)

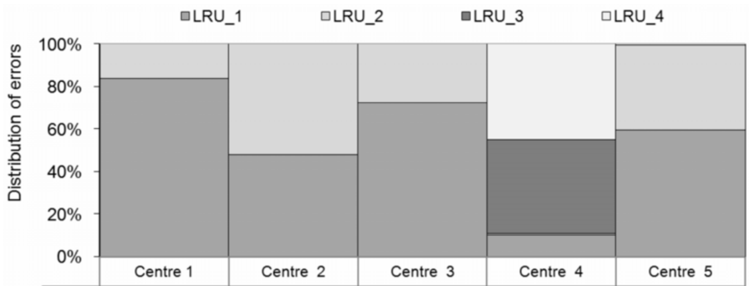


	Centre 1	Centre 2	Centre 3	Centre 4	Centre 5
Bui_6	97.1%	16.8%	41.7%	14.1%	59.1%
Bui_5			41.7%		
Bui_4			2.8%		6.2%
Bui_3		82.8%	5.6%	2.3%	2.6%
Bui_2			2.8%		0.6%
Bui_1	2.9%		5.6%	83.7%	31.4%

Figure 7. Distribution of errors in buildings (descriptive data)

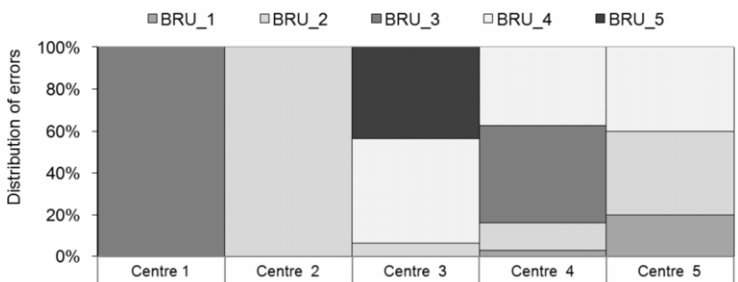
	Centre 1	Centre 2	Centre 3	Centre 4	Centre 5
Apa_3			100.0%	17.2%	
Apa_2	100.0%	100.0%		29.6%	
Apa_1				53.2%	100.0%

Figure 8. Distribution of errors in apartments (descriptive data)



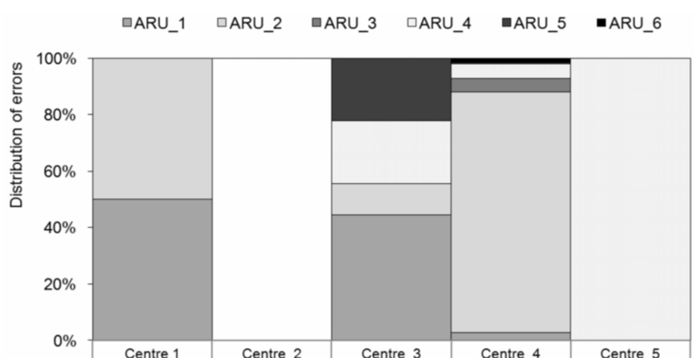
	Centre 1	Centre 2	Centre 3	Centre 4	Centre 5
LRU_4	0.0%	0.0%	0.0%	44.6%	0.6%
LRU_3	0.0%	0.0%	0.0%	44.0%	0.0%
LRU_2	16.0%	51.9%	27.8%	0.8%	40.0%
LRU_1	84.0%	48.2%	72.2%	10.0%	59.4%

Figure 9. Distribution of errors in land registration units (descriptive data)



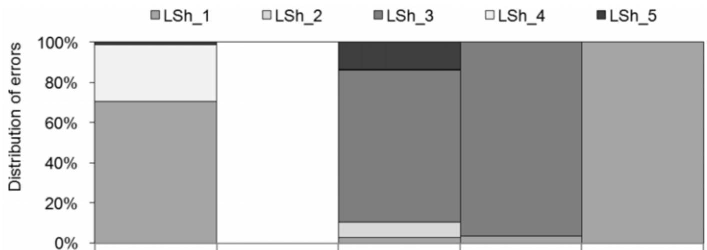
	Centre 1	Centre 2	Centre 3	Centre 4	Centre 5
BRU_5			43.7%		
BRU_4			50.0%	36.8%	40.0%
BRU_3	100.0%			46.1%	
BRU_2		100.0%	6.3%	13.2%	40.0%
BRU_1				2.6%	20.0%

Figure 10. Distribution of errors in building registration units (descriptive data)



	Centre 1	Centre 2	Centre 3	Centre 4	Centre 5
ARU_6				1.6%	
ARU_5			22.2%	0.5%	
ARU_4			22.2%	5.2%	100.0%
ARU_3				4.7%	
ARU_2	50.0%		11.1%	85.5%	
ARU_1	50.0%		44.4%	2.6%	

Figure 11. Distribution of errors in apartment registration units (descriptive data)

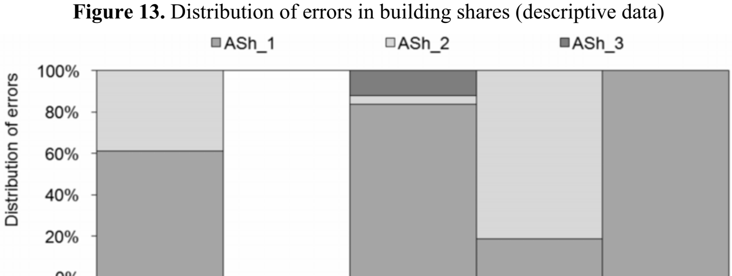


	Centre 1	Centre 2	Centre 3	Centre 4	Centre 5
LSh_5	1.4%		13.3%		
LSh_4	28.1%		0.3%		
LSh_3			75.9%	96.6%	
LSh_2			7.7%		
LSh_1	70.5%		2.6%	3.4%	100.0%

Figure 12. Distribution of errors in land shares (descriptive data)

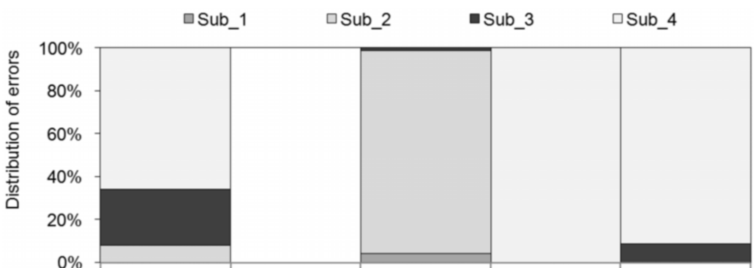
	Centre 1	Centre 2	Centre 3	Centre 4	Centre 5
BSh_5			35.6%		
BSh_4			0.5%		
BSh_3			1.4%		
BSh_2			19.4%		
BSh_1	100.0%		42.7%		

Figure 13. Distribution of errors in building shares (descriptive data)



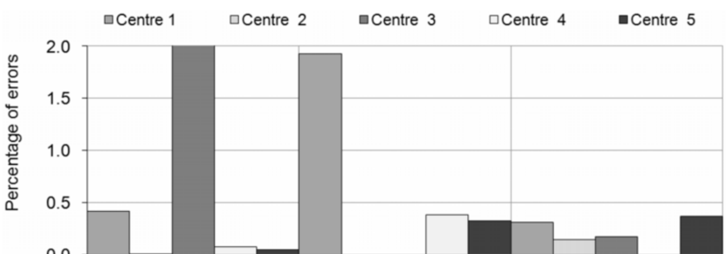
	Centre 1	Centre 2	Centre 3	Centre 4	Centre 5
ASh_3			12.0%		
ASh_2	38.7%		4.0%	81.2%	
ASh_1	61.3%		84.0%	18.8%	100.0%

Figure 14. Distribution of errors in apartment shares (descriptive data)



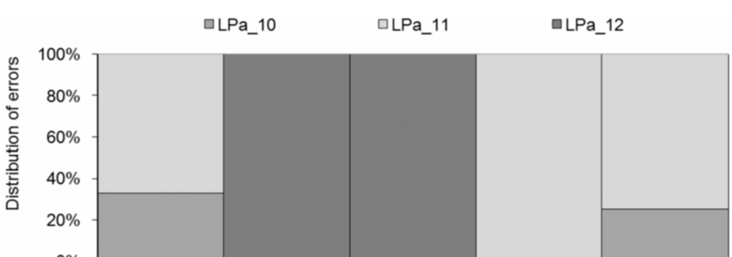
	Centre 1	Centre 2	Centre 3	Centre 4	Centre 5
Sub_4	65.9%			100.0%	91.4%
Sub_3	26.1%		1.1%		8.6%
Sub_2	8.0%		94.1%		
Sub_1			4.3%		

Figure 15. Distribution of errors in cadastral subjects (descriptive data)



	LPa	Bui	LUs
Centre 1	0.42	1.93	0.31
Centre 2	0.01		0.15
Centre 3	25.80		0.17
Centre 4	0.08	0.38	
Centre 5	0.05	0.33	0.37

Figure 16. Percentage of errors in geometric data



	Centre 1	Centre 2	Centre 3	Centre 4	Centre 5
LPa_12		100.0%	100.0%		
LPa_11	67.1%			100.0%	75.0%
LPa_10	32.9%				25.0%

Figure 17. Distribution of errors in land parcels (geometric data)

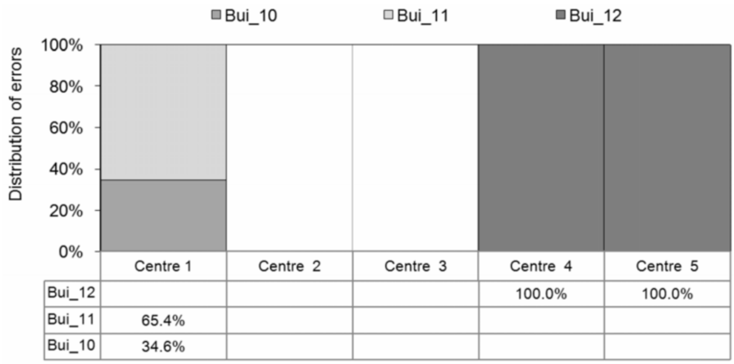


Figure 18. Distribution of errors in buildings (geometric data)

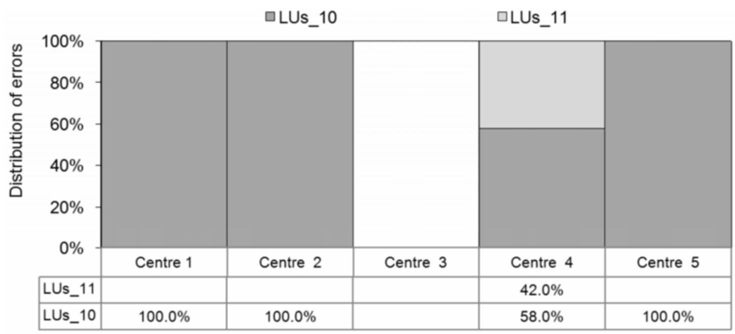


Figure 19. Distribution of errors in land uses (geometric data)

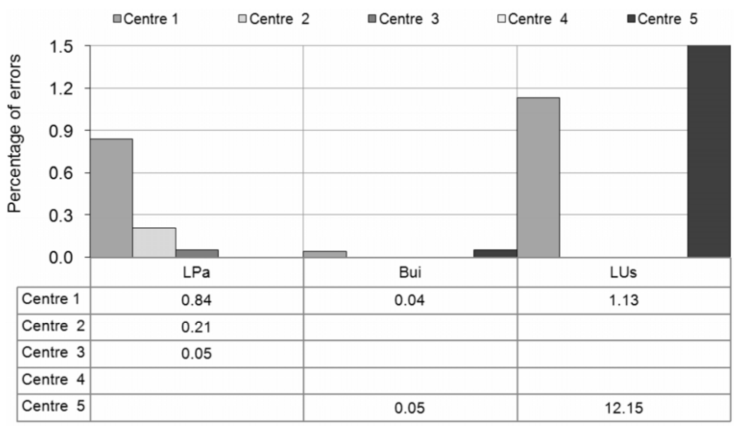


Figure 20. Percentage of topological errors in geometric data

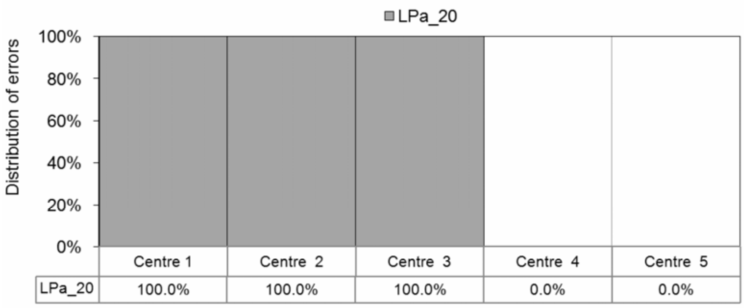


Figure 21. Distribution of errors in land parcels (topological data)

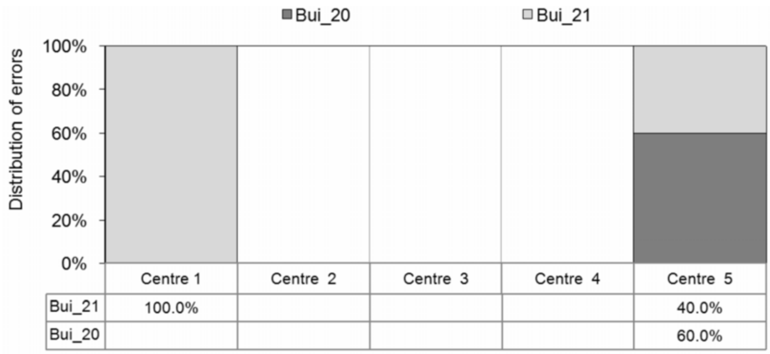


Figure 22. Distribution of errors in buildings (topological data)

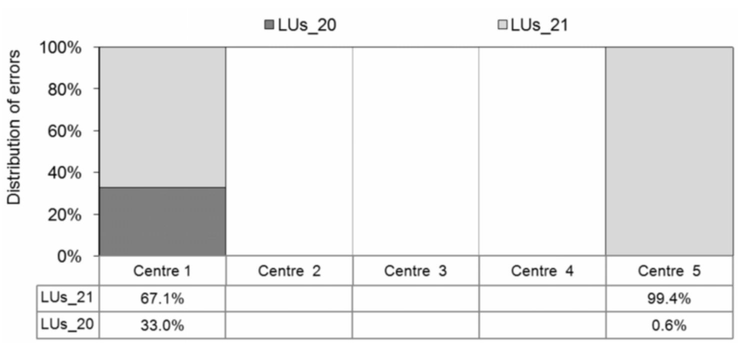


Figure 23. Distribution of errors in land uses (topological data)

In Figure 20 the percentage of cadastral objects with topological errors found in geometric data is presented.

In Figure 21 the distribution of topological errors in land parcels is shown. Following denotation of errors is used: LPa_20 – lacking polygons.

In Figure 22 the distribution of topological errors in buildings is shown. Following denotation of errors is used: Bui_20 – lacking polygons; Bui_21 – overlapping polygons.

In Figure 23 the distribution of topological errors in land uses is shown. Following denotation of errors is used: LUs_20 – overlapping polygons; LUs_21 – lacking polygons.

3. Conclusions

There are many reasons for insufficient quality of data in legacy cadastral systems. First of all the development of information technology caused multiple data transfers from older to newer systems, many times without adequate control of correctness. Next, the regulations concerning the cadastre were changed many times in Poland, on average every five years. Insufficient funds caused that information centres were poorly equipped and there was high staffing fluctuation. The quality of cadastral system application themselves had also substantial impact on the level of lacks and defects in data. The applications lack appropriate procedures to check data during input and to monitor quality of databases continuously. Sometimes data were input into cadastral systems by outsourcing geodesic companies. Moreover, there was insufficient quality control during data preparation and final loading into production databases. On the other hand, each year information centres were obliged to export data to the central IACS system using the SWDE format. In order to succeed, the centres were forced to control and repair data to be transferred. We may hope that modern integrated cadastral systems, like the Kataster OnLine, based on the newest information technology and unified object approach, will be leading step by step to the improvement of cadastral data.

References

[1] M. Bobrowski, M. Marre, D. Yankelevich: *Measuring Data Quality*, Universidad de Buenos Aires. Report 99-002, Buenos Aires, Argentina, (1999).

[2] Data Monitoring: *Taking Control of Your Information Assets*, DataFlux Corp., (2004).

[3] L. English: *Improving Data Warehouse and Business Information Quality*. Wiley, (1999).

[4] H. Hinrichs, T. Aden: *An ISO 9001:2000 Compliant Quality Management System for Data Integration in Data Warehouse Systems*, Proceedings of the International Workshop on Design and Management of Data Warehouses, Interlaken, Switzerland, (2001).

[5] D. Król, T. Lasota, M. Siarkowski, B. Trawiński: Investigation of Application Specific Metrics to Data Quality Assessment, *Lecture Notes in Computer Science*, 4439, (2007), 438-448.

[6] Król D., Lasota T., Svoboda L., Trawiński B., Widz R.: Kataster OnLine - modern integrated cadastral information system, in: *Multimedia and network information systems*. Proceedings ed. by A. Zgrzywa. Wrocław, September 21-22, 2006. Wrocław: Oficyna Wydaw. PWroc., (2006), 371-383 .

[7] D. Król, B. Trawiński, W. Zawiła: Integration Techniques and Approaches to Cadastral and Financial-Accounting Systems, *International Journal of Intelligent Information and Database Systems*, 2008 (in print).

[8] D. Król, B. Trawiński, W. Zawiła: Problems of the integration of cadastre information systems. In: *Multimedia and network information systems*, Proceedings ed. by A. Zgrzywa. Wrocław, (2006), 237-246.

[9] Y. W. Lee, L. L. Pipino, R. Y. Wang: Data Quality Assessment. *Communications of the ACM* 45 (2002), 211-218.

[10] Y. W. Lee, D. M. Strong, Wang R. Y.: Data Quality In Context, *Communications of the ACM* 40 (1997), 103-110.

[11] P. Sieniawski, B. Trawiński: An open platform of data quality monitoring for ERP information systems. In: *Software engineering techniques: design for quality*, Ed. by K. Sacha, Boston: Springer, (2006), 289-299.

This page intentionally left blank

Web Systems and Network Technologies

This page intentionally left blank

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-243

Web-Based Recommender Systems and User Needs – the Comprehensive View

Przemysław KAZIENKO
Technical University of Wrocław, Institute for Applied Informatics
Wybrzeże Wyspiańskiego 27, 53-370 Wrocław, Poland
e-mail: przemyslaw.kazienko@pwr.wroc.pl

Abstract. Recommender systems became an important element of many web-based systems. The general process about how recommender systems fulfil various user needs based on the analysis of recent literature in presented in the survey paper. Recommender systems are compared with other method like Information Retrieval and browsing. The paper also includes the comprehensive view over the typical components and structure of recommender systems. The state-of-the-art of recommendation methods is described as well.

Keywords. recommender system, Information Retrieval, survey, user needs

Introduction

Recommender systems are now an integral part of many web sites, including e-commerce (Amazon, CDNow), news portals (ACR News, INFOrmer), travel portals (LifeStyle Finder) and many, many others [76]. Moreover, recommender systems can provide an effective form of targeted marketing by creating a personalized shopping experience for each e-commerce customer [46].

The main goal of a recommender system is to provide more or less personalized list of objects relevant for the active user. These objects should possibly meet current needs of the user. The year by year increasing popularity of recommendations has made visitors more used to utilize recommender systems incorporated into web sites they visit. Furthermore, e-commerce users are highly influenced by recommender systems [88] since "personalized product recommendations shown to strongly influence consumer shopping behaviour" [20]. The users simply want recommendations "Consumers indicate a strong preference for sites that provide personalized product recommendations, with 45 percent claiming that they are more likely to shop at sites with personalized recommendations than at sites without them" [20]. However, users more and more expect services at the high level and "39 percent of consumers are less willing to return to sites that provide poor quality recommendations and 35 percent are less willing to buy products from those sites." [20].

1. User Needs

Users of information systems have their own needs these systems can fulfil by means of specialized mechanisms like browsing, searching or recommendations. They enable to select the objects corresponding to user needs (Figure 1).

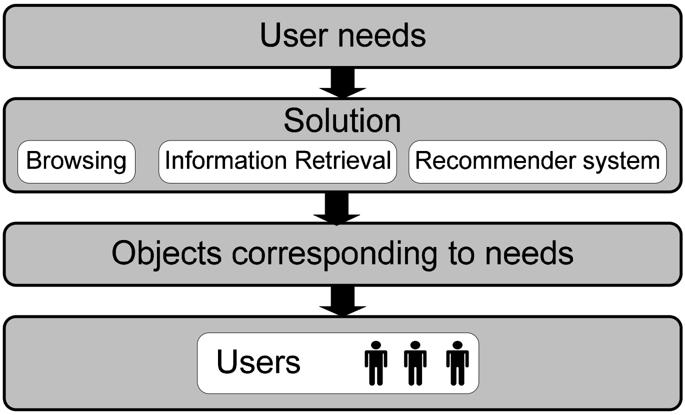


Figure 1. Fulfulling user needs in information systems

Based on the literature analysis, the typical user needs are (Figure 2):

1. Shopping. Products or services to order or to buy [1], [8], [31], [44], [70], [102], [112], [113], e.g. specialized financial services [29], travel destinations [30], [84], [89], planning tours in the city [12], mobile applications to install [106], web services to use [7], [103], e government services to utilize [34], selection of only trustworthy sellers in the auction system [77]
2. Tasks to do [57], [79], especially next task to do or to learn [71]
3. Contents to read or watch, especially textual or multimedia, for example: scientific articles [101], Music [14], [26], [68], movies [13], [21], [22], TV shows [111]
4. Navigation in the web sites [47], [48], for example in the school web site [64].
5. Learning [24], [96], in particular in distributed environments [23], [25] that also include mobile solutions [3], [4]. Users can be assisted while training to utilize office suits [71]. The needs of learners usually comprise the help at decision what to do (learn) next, i.e. the best next learning activities. For lifelong learners it was considered in [27], [39], English lessons for learners for which English as a second language – ESL students [37], analysis of architectural precedents similar to the currently studied case studies for architectural students [81].
6. Human relationships [10], [42], [51], [94], willingness to meet other people who visit the same places and in this way belong to the same informal group [35].
7. Acceptable quality of service (QoS) for content delivery; important especially in case of multimedia data [98], [110]. It also includes user interface (UI) adaptation [38]
8. Personalized and targeted advertising [5], [40], [46], [59], [60], [83], [108] also for unsought products [69]
9. Management of web sites. It includes recommendation of hyperlinks that need to be removed or promoted [45], [53], navigational paths that require simplifications [93] or entire pages that necessitate corrections [92].

10. More specific needs include: product and services that should be designed or developed [75], conducting negotiations [66], recipes for cooking [90], profit increasing in e-commerce [16], production of multi-channel TV shows [111], context-aware information providing for drivers [105], alignment of ontologies [97].

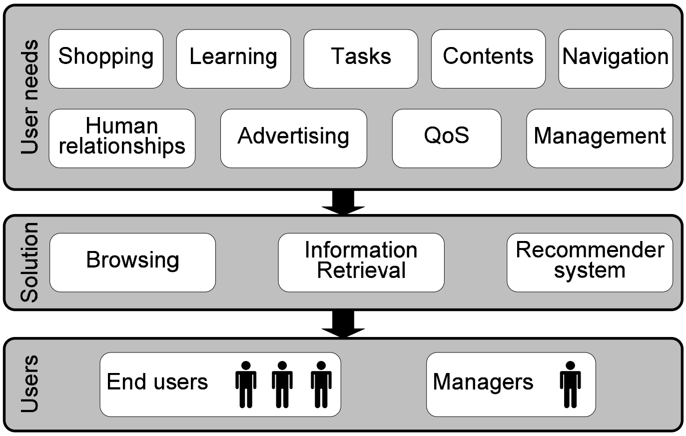


Figure 2. Recommender system as one of the solutions meeting user needs

2. Solutions Fulfilling User Needs

To meet the above needs a system usually provides diverse methods that enable the extract the objects suitable for the users. Three main approaches can be distinguished: simple browsing, Information Retrieval (IR) and recommender systems (Fig. 2). All they try to matches user preferences with the objects available in the system.

The first method, browsing consists in using navigational tools to move from one item to another in order to find the proper one. It is inefficient but very popular way. The two next approaches are in a sense similar each other. The main difference is that Information Retrieval utilizes query-response model. In the raw model of IR, the users provide their queries using some specialized interfaces. However, this appears to be impractical in case of diverse and large environments e.g. huge online retailer that can contain tens of millions of customers and millions of distinct catalogue items like Amazon.com [70]. In such case, formulation of precise queries can be tricky for the user. Hence, the query should to be routed not to the entire database but only to its selected subset or summary of the data. This in turn leads to reduction of output quality. Besides, it is inconvenient for the user to provide information about their recent needs. Moreover, it may be hard for them to express these needs in the precise way. This is probably the main disadvantage of Information Retrieval idea. Additionally, queries in IR are typically processed online so they may be inefficient in the large scale environments.

Nevertheless, in some solutions, Information Retrieval and recommender systems can be close each other, e.g. while recommending “other objects like this one” [71]. On the other hand, in Information Retrieval, users are able to express their needs in the direct way.

Concluding, query-based methods appear to be unfeasible in large online solutions that operate on thousands of objects of different kind or structure.

3. Solutions Fulfilling User Needs

To meet the above needs a system usually provides diverse methods that enable the extract the objects suitable for the users. Three main approaches can be distinguished: simple browsing, Information Retrieval (IR) and recommender systems (Fig. 2). All they try to matches user preferences with the objects available in the system.

Table 1. Comparison of general concepts used to deliver the right objects and fulfil user needs

Feature	Browsing	Information Retrieval	Recommender systems
Ease of use	Depends on solution	Low [1]	High
Relevance to user needs	Low-medium[2]	High	Low-medium [3]
Completeness of results	Small/medium	High [4]	Small
User effort to find the right objects	Large	Medium	No or little effort [5]
Speed of access to the most suitable objects	Small	High	Medium
Expression of recent needs	Direct	Direct	Indirect[6]
Load of online processing	None [7]	Large in case of many queries	Small
Load of offline processing	Small [8]	Small [9]	Large (DM methods)
Maintenance of additional data	No	Some[10]	Yes[11]

[1] Especially in case of less structuralized objects or objects many attributes
[2] It can be difficult to find the relevant contents in case of larger systems
[3] Depends on personalization level
[4] Compared to browsing and recommendations
[5] Some effort are required to maintain user profiles or preferences in some solutions
[6] Users can express their general preferences in some recommender systems. In other solutions, the system monitors user activities and treats them as hints for recommendations – indirectly provided current needs. In item-to-item (content-based) recommendations, the system matches the recently viewed objects with the other, relevant ones. The user need is then expressed by the recently viewed object
[7] Some online activities are required in case of dynamically generated contents.
[8] Maintenance of navigation paths, e.g. hyperlinks
[9] Offline maintained indexes and calculations of ranking functions e.g. PageRank in Google
[10] It refers auxiliary data used to speed up responses, e.g. indexes
[11] It refers specific data sources used for recommendation, see sec. 6

The first method, browsing consists in using navigational tools to move from one item to another in order to find the proper one. It is inefficient but very popular way. The two next approaches are in a sense similar each other. The main difference is that Information Retrieval utilizes query-response model. In the raw model of IR, the users provide their queries using some specialized interfaces. However, this appears to be impractical in case of diverse and large environments e.g. huge online retailer that can contain tens of millions of customers and millions of distinct catalogue items like Amazon.com [70]. In such case, formulation of precise queries can be tricky for the user. Hence, the query should to be routed not to the entire database but only to its selected subset or summary of the data. This in turn leads to reduction of output quality. Besides, it is inconvenient for the user to provide information about their recent needs. Moreover, it may be hard for them to express these needs in the precise way. This is probably the main disadvantage of Information Retrieval idea. Additionally, queries in IR are typically processed online so they may be inefficient in the large scale environments.

Nevertheless, in some solutions, Information Retrieval and recommender systems can be close each other, e.g. while recommending "other objects like this one" [71]. On the other hand, in Information Retrieval, users are able to express their needs in the direct way.

Concluding, query-based methods appear to be unfeasible in large online solutions that operate on thousands of objects of different kind or structure.

Table 1 contains the comprehensive view of all three methods used to meet user needs, as well as their profiles.

4. Users of Recommender Systems

There are two main types of recommender system users: end users and web site managers. The former are those who are suggested more or less personalized objects that fulfil their current or general needs. The latter wants to influence the recommendation process to offer to the users objects that are potentially interesting for the users and simultaneously satisfy some business requirements, e.g. the more profitable goods [16], unsought products [69]. Besides, manages may wish to promote objects that are most likely to be followed by certain user actions after their presentation, e.g. advertisements most likely to be clicked [46], products most likely to be purchased, or bids most likely to finish negotiations [66]. Sometimes manages are willing to extract and endorse other users that may be attractive or trustworthy for their clients [77]. There is also web content which should be suggested with the most or the least popular objects. The former can be in turn highlighted while the latter removed. It can refer usability assessment for hyperlinks maintained on web pages according to their usage by users [53].

5. Environment and Structure of Recommender Systems

A recommender module operates in the environment of the system it is incorporated in. Using various data sources (see sec. 6) available in the system; the recommender agent tries to discover useful patters existing in the data (Fig. 3). Different statistical or data mining methods can be utilized for this purpose (see sec. 8).

Having the patterns extracted, the recommender system makes use of one or more recommendation techniques to rank the selected objects. Afterwards, these objects are presented to the user (Fig 4). In the selection process, information about current or general user preferences are exploited to facilitate the more personalized recommendation.

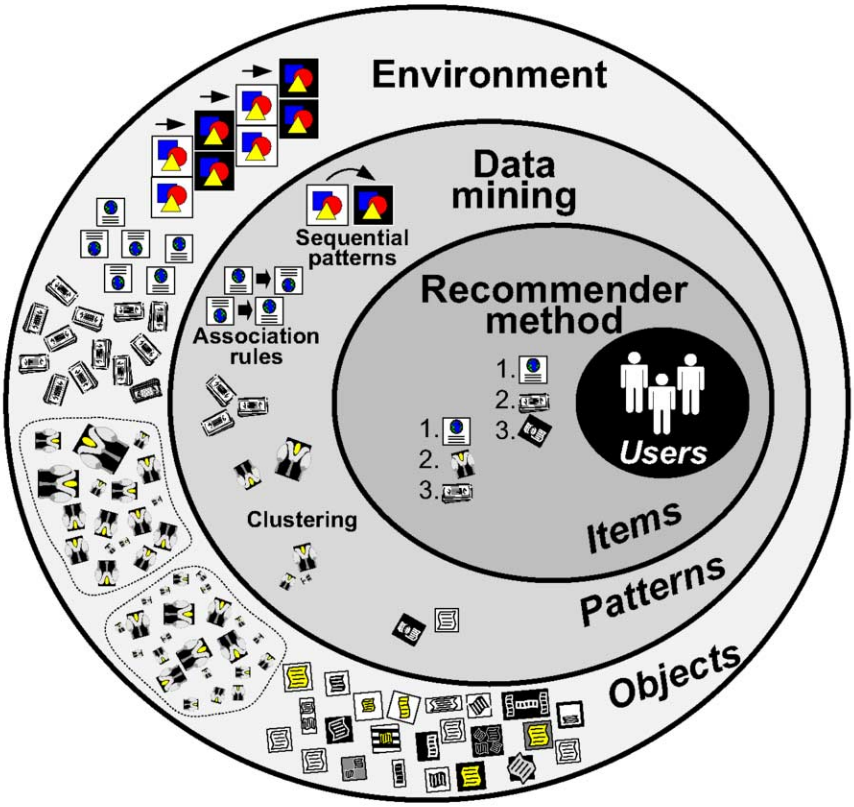


Figure 3. The environment of recommender systems

6. Data Sources

The recommender system can utilize many different data sources available, in particular the following ones: user profiles, content and usage data, structure of the web site, user ratings, and data about communication between users as well as visible or invisible relationships between users (Fig. 4). The problem of user general profile and preferences maintenance has been studied in science for many years, especially in the field of Information Retrieval. In recommender systems user profiles are also widely used [2], [76], [81], [86], [87]. They can have a form of WordNet-based textual user profiles [22]. Attributes of products or services that are utilized as data sources can be more or less specific and include their critiques [15], their filling the shelves [69], textual reviews [1], [113], timetables of tours [12], synsets (synonym sets) tokenized

from the textual contents and processed owing to the WordNet knowledge database [22] as well as textual descriptions of multimedia objects, e.g. movies [22].

Usage data in case of web sites usually refer to web server logs [32], [48], [43], [45] also limited to only related to clicks while placing items into the basket [19].

Structure data refers information about direct relationships between objects. In web systems, this structure is expressed by hyperlinks incorporated in web pages [53], [67].

Another typical data source used by recommender systems are ratings provided by users [22], [25]. It can also be ratings about other users (their photo) in online dating system [10], pseudo rating data derived from the implicit feedback data [65] or even group ratings [17].

Communication between users as well as either visible or invisible relationships linking users can also be utilized for recommendation [51], [109]. They reflect direct similarity between humans that utilize the system. Hence, they can replace typical collaborative filtering approach (see sec. 8.4). Visible and invisible relationships between users exist in the systems in which users can maintain their contacts to others or meet one another at different activities [42].

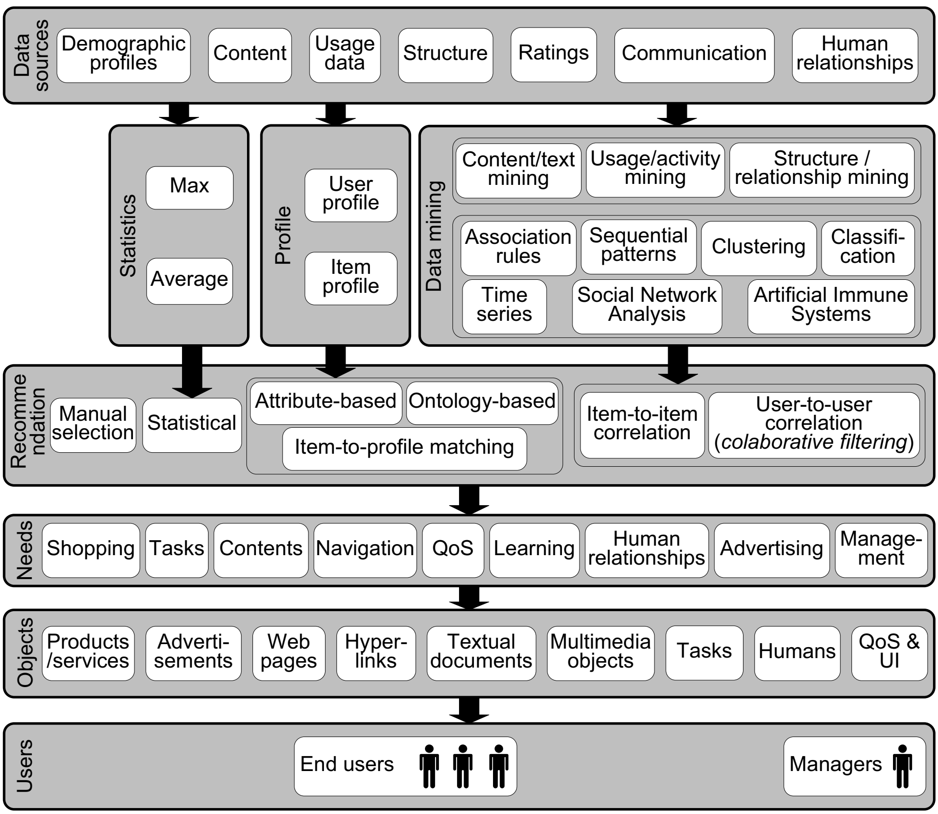


Figure 4. The structure of recommender systems

More specific data source cover: geographical position that facilitates context-aware recommendation of interesting places for mobile tourists [89], localization that identifies informal groups of people [35], learning performances [25], UDDI catalogues [7, 103], user preferences related to the objects specified in the explicit way [75],

business transactions [77], semantic content annotations to TV programs [111], WordNet database [22], and data included in anthologies [52, 97].

7. Recommended Objects

Recommended objects are suggested to the users in order to fulfil their needs. Their profile depends on the systems but the most typical are products [1], [15], [16], [17], [31], [70], [112], [113] or services of different kind, among others: financial services [29], travels [30, 84], tours in the city [12], applications to be installed on mobile devices [106], web services [7], [103], e-government services [34], groups (families) of products or services that should be developed [75], unsought products [69], some specific products like cosmetics [102]. The related objects are web advertisements, usually in the form of banners [46].

Other typical kind of recommended items in the web environment are hyperlinks [9], [47] and web pages [48], e.g. school web sites [64].

The third group of objects that can be suggested to the users are textual documents, e.g. in the digital library [101] or multimedia objects like music [14], [26], [68], TV shows [111], TV news [41] and TV programs [6], [80], [107], movies [13], [21], [22], and even architectural objects [81].

Contents or tasks especially in e-learning [23], [24], [25], [27] constitute yet another type of recommended objects. It also includes the best next learning activities [39], entire courses at the university [28], recommendation of scientific and technical papers (documents) related to the curriculum provided by tutors [96], reading English lessons for ESL (English as a second language) students [37], architectural objects related to the cases analyzed by students [81], etc. However, tasks can also refer to other domains, in particular to “tasks to do”, e.g. in a cadastre information system [57] or next functions in office suits [71], [79].

The emerging and relatively new kind of recommendations are recommendation of humans, including self recommendations [73], suggestions in dedicated (online) social network systems [42], [51], recommendations in dating systems [10], matching humans or groups of humans visiting the same places [35], and even discovery of trusty and reliable sellers [77].

The appropriate parameters of content delivery, i.e. QoS (Quality of Service) [98], especially in pervasive learning solutions [110], may be suggested to the users according to their profiles and preferences.

There are also solutions that operate on some more specific objects for recommendation, such as offers in negotiations [66], points of interest (POI) for mobile tourists [89], architectural precedents [81], products matching orders in the ERP system [95], news [63] including financial news that match current market behaviours [61], jokes [33], [78], flights [104], cooking recipes [91], gas stations [105], places where the informal groups the user belongs meet one another [35] and strategies for ontology alignment [97].

The above list of types of recommended objects is not an in-depth catalogues. There are many systems, also working commercial ones that suggest to their users subjects not mentioned above.

8. Recommendation Methods

The recommender system utilizes data sources (see sec. 6) to rank objects (see sec. 7) for recommendation (Fig. 4). Three main groups of methods, which provide initial patterns useful for recommendation, can be distinguished: statistics, profile maintenance and data mining (see sec. 8.3).

At the next stage, the patterns are used to create final recommendation lists. There are five basic approaches for this purpose (Fig. 4): manual selection, statistical methods (see sec. 8.1), profile matching – stereotyping, i.e. item-to-profile matching (see sec. 8.2), item-to-item correlation (content-based filtering) and user-to-user correlation called collaborative filtering (see sec. 8.4).

8.1. Statistical Methods

Since the data sources usually contain vast amounts of data, there is a necessity to use some efficient algorithms and techniques to process these data sets. The usual solution of this problem is usage of data mining methods. They are able to manage with plenty of data records in the natural way. However, there are approaches that make use of simplified techniques. For example, statistical maximum or average values can be easily calculated even for large data sets. Moreover, in practice, they are used in many web services in the form of "best buy" or "top rated" or the highest frequency among other users, e.g. at learning to become skilled users of the word processor [71]. According to experiments in [49], these statistical approaches turn out to be especially useful at the start of the system – they simply solve the so-called "cold start problem". However, afterwards, the other methods like association rules (data mining) appear to provide more sophisticated outcome that is handier for users. Besides, statistic-based suggestions like "best buy" do not distinguish users. All users obtain the same recommendations regardless their needs or preferences.

8.2. Profile Matching

There are also approaches that try to match user and object profiles. These profiles need to contain equivalent attributes. For example, a user who can speak Polish is suggested movies with Polish dubbing or at least Polish subtitles. The systems must understand the item domain very well. It must have knowledge of important features of the item, and be able to access the knowledge base where these important features are stored in an inferable way.

The main disadvantages of this solution are necessity of profile maintenance and update as well as troubles in the environments with hardly structuralized objects or objects of many different structures – the separate sets of attributes for each data structure. Profile matching is item-to-profile recommendation sometimes also called knowledge-based recommendation [99] or demographic filtering [58, 82].

8.3. Data Mining

For all the disadvantages mentioned in sec. 8.1 and 8.2, data mining methods are widely used in more advanced recommender systems.

Plenty of different data mining methods can be applied to recommender systems. They include association rules [72] that can be integrated with clustering [62] and

applied to advertising [59]. Indirect association rules can significantly extend recommendation list [44] while negative association rules facilitate to reorder recommendation lists built upon content similarity between objects [43, 45, 53]. Association rules are extracted from the source data set that contains information about co-occurrences of objects. In case of web environment the appropriate information are stored in web logs [49, 64] or customer purchases or baskets [19].

Other methods, in a sense similar to association rules, are sequential patterns. In opposite to association rules, which operate on unordered sets, sequential patterns deal with sequence that respects the order over time. Many recommender systems were developed based on sequential pattern analysis, e.g. [19], [32], [54], [114]. The novel type of sequential patterns are sequential patterns with negative conclusions that similarly to negative association rules enable negative verification of previously created recommendation lists [43], [45].

In data mining association rules, sequential patterns and time series (similar times sequences) are described in one common term link analysis. In recommender systems, time series and trends, which have been discovered within them, can help to predict the influence of future financial news on the market [61].

Even more popular method of data mining utilized in recommender systems is clustering. Its main concept consist in extraction of small number of groups which component members are similar one another within the cluster and dissimilar to members from other groups. Additionally, each cluster can have a representative that is used for comparisons with the considered object related to the current user. Hence, clustering can be applied to users or customers themselves [3], [71], [72], including agglomerative clustering of users [36]. The typical and one of the simplest clustering algorithms k-means can be enhanced with genetic algorithms for initial seed generation [55]. Users may be clustered together with places they meet one another [35] and combined with reputation the sellers possess [77]. Also user profiles extracted from the textual content can be clustered [22]. Yet another application is clustering of historical web user sessions [47] or other usage data like clicked advertisements [46] and even web textual contents [48]. Clustering is the common method utilized in user-to-user or item-to-item correlation method (see sec. 8.4).

Other data mining methods useful in recommender systems are: classification including combination (associative) of various classification methods [112], prediction of ratings based on Bayesian model [100], also with manual interference of users [85], the naïve Bayes inductive learning method used for user profiling [22], support vector machine method (SVM) [18], also boosting-SVM [69], artificial immune system (AIS) [13, 91], social network analysis (SNA) [42], [51], [50], [52], [109], case-based reasoning (CBR) [12].

Data mining method can be applied either to web content (text mining) [1], [43], [47], [48], [46], [113], web usage (usage mining) i.e. usually web logs or to web structure (web structure mining).

8.4. Content-based Methods, Collaborative Filteringand Hybrid Approaches

Data mining methods deliver useful patterns that can be utilized to create recommendation list which are next presented to the user. There are two main approaches that make use of patterns delivered by data mining methods: the content-based item-to-item correlation [11], [82] and user-to-user correlation usually called collaborative filtering. There exist also many hybrid solutions

The former tries to find objects that are most similar or appropriate for the just viewed one. This correspond to ephemeral personalization [49], [86], i.e. the system

delivers a different list on every page of the website but be the same for all users. Hence, personalization depends on the item (web page) which is recently viewed by the current active user. Recommendation lists can be calculated upon the source data (see sec. 6) using some data mining methods that finds correlations between items. It refers especially association rules [64], sequential patterns and less frequently – clustering. It can also be achieved by using quite simple feature similarity function, e.g. features of popular music [26] or ontology-based similarity [52], [97]. Item-to-item similarity measure may have the form of Pearson-r correlation coefficient [81]. Moreover, content-based filtering can make use of deeper knowledge about recommendation domain, e.g. financial services. Thus, it may include the interaction with the user [29].

Collaborative filtering facilitates to obtain persistent personalization [49, 86], in which users are suggested items according to their general needs – the system generates a different recommendation list for each user. Obviously, it can only work with identified, logged in users. Collaborative filtering [10], [13], [17], [65], [74], [96] reflect user-to-user-correlation and sometimes is also called social-based [39]. Typical collaborative approaches compute a similarity between the current user and all other users by taking into account their ratings, i.e. the set of ratings provided on the same items. Based on these ratings of the most similar users, commonly referred to as neighbours, collaborative algorithms calculate recommendations separately for each current user. In case of large systems that maintain thousands of users, such online evaluation can be very tricky. For that reason, some data mining methods are used, in particular clustering [47], [46]. It enables computing the similarity only between the current user and the entire neighbourhoods, i.e. representatives of each cluster. One nearest neighbourhood forms the set of users closest to the current one. The problem with collaborative filtering is that the similarity values are only computable if users have some common rated items. There are also "cold start" (lack ratings for the new user or the new item) and sparse problem (only few ratings for each users in the environment of large number of items).

These significant shortcomings can be usually overcome by the hybrid recommender systems [11] that usually extend collaborative filtering with content-based approaches [3], [34], [37], [46], [47], [48], [81], [101], [102], [106]. Item-to-item similarity can be combined with the seller-to-seller trust-based clustering. As a result, auction bids that both are related to the just viewed product and are offered by reliable sellers are recommended [77]. In another hybrid approach, Separate methods for different classes of objects, are used, e.g. one method is used for museums and another for restaurants in the tourist recommender system [89]. Combination of collaborative filtering and clustered content-based user profiles which are automatically extracted and maintained by means of textual analysis of content as well as lexical knowledge stored in the WordNet database were presented in [22]. Usage of collaborative filtering for old users and item-to-user approach for new ones was proposed in [99]. Sobecki used fuzzy inference for demographic stereotype reasoning [91].

9. Conclusions and Future Work

Recommender systems have been studied in the scientific literature for about ten years. As a result many specific and versatile systems have been developed. All of them try to meet diverse user needs and constitute alternate concept in relation to Information Retrieval and simple browsing. Their main advantage is usually minimal amount of

effort the user needs to undertake – the system performs all necessary tasks itself based on different data sources including user profiles or activities. Recommender systems can utilize various concepts like simple statistics, profile matching (item-to-user correlation), and a variety of data mining techniques, content-based methods (item-to-item correlation), collaborative filtering (user-to-user correlation) and a range of hybrid approaches.

Future work will focus on emerging methods of recommendation, in particular negative recommendation and social-based those make use of social network analysis (SNA).

Acknowledgments.

The work was supported by The Polish Ministry of Science and Higher Education, grant no. N516 037 31/3708.

References

[1] Aciar S., Zhang D., Simoff S.J., Debenham J.K.: Recommender System Based on Consumer Product Reviews. WI 2006, IEEE Computer Society, 2006, 719-723.
[2] Adomavicius G., Tuzhilin A.: Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Transactions on Knowledge and Data Engineering 17 (6), 2005, 734-749.
[3] Andronico A., Carbonaro A., Casadei G., Colazzo L., Molinari A., Ronchetti M.: Integrating a multi-agent recommendation system into a Mobile Learning Management System. AIMS 2003, http://ai-gate.cs.uni-sb.de/%7Ekrueger/aims2003/camera-ready/carbonaro-4.pdf.
[4] Andronico A., Carbonaro A., Colazzo L., Molinari A.: Personalisation services for learning management systems in mobile settings. International Journal of Continuing Engineering Education and Life Long Learning 14 (4-5), 2004, 353-369.
[5] Bae S.M., Park S.C., Ha S.H.: Fuzzy Web Ad Selector Based on Web Usage Mining. IEEE Intelligent Systems 18 (6), 2003, 62-69.
[6] Baudisch P., Brueckner L.: TV Scout: Lowering the Entry Barrier to Personalized TV Program Recommendation. From Integrated Publication and Information Systems to Virtual Information and Knowledge Environments, Essays Dedicated to Erich J. Neuhold on the Occasion of His 65th Birthday. LNCS 3379, Springer Verlag, 2005, 299-309.
[7] Blake M.B., Nowlan M. F.: A Web Service Recommender System Using Enhanced Syntactical Matching. ICWS 2007, IEEE Computer Society, 2007, 575-582.
[8] Bhasker B., Park H.-H., Park J., Kim H.-S.: Product Recommendations for Cross-Selling in Electronic Business. AI 2006, LNAI 4304, Springer Verlag, 2006, 1042-1047.
[9] Bollen J.: Group User Models for Personalized Hyperlink Recommendations. AH 2000, LNCS 1892, Springer Verlag, 2000, 38-50.
[10] Brozovsky L., Petrıcek V.: Recommender System for Online Dating Service. Znalosti 2007, also Computer Research Repository (CoRR), cs/0703042, March 2007, http://www.occamslab.com/petricek/papers/dating/brozovsky07recommender.pdf.
[11] Burke R.: Hybrid recommender systems: survey and experiments. User Modeling and User-Adapted Interaction 12(4), 2002, 331-370.
[12] Castillo L., Armengol E., Onaindía, E., Sebastiá L., González-Boticario J., Rodríguez A., Fernández S., Arias J.D., Borrajo D.: samap: An user-oriented adaptive system for planning tourist visits. Expert Systems with Applications 34 (2), 2008, 1318-1332.
[13] Cayzer S., Aickelin U.: A Recommender System based on the Immune Network. CEC2002, 2002, 807-813.
[14] Chen H.-C., Chen A. L. P.: A Music Recommendation System Based on Music Data Grouping and User Interests. ACM CIKM, ACM, 2001, 231-238.
[15] Chen L., Pu P.: The evaluation of a hybrid critiquing system with preference-based recommendations organization. RecSys 2007, ACM, 2007, 169-172.

[16] Chen L.-S., Hsu F.-H., Chen M.-C., Hsu Y.-C.: Developing recommender systems with the consideration of product profitability for sellers. Inf. Sciences 178 (4), 2008, 1032-1048.
[17] Chen Y.-L., Cheng L.-C., Chuang C.-N.: A group recommendation system with consideration of interactions among group members. Expert Systems with App. 34 (3), 2008, 2082-2090.
[18] Cheung K.-W., Kwok J.T., Law M.H., Tsui K.-C.: Mining customer product ratings for personalized marketing. Decision Support Systems 35 (2), May 2003, 231-243.
[19] Cho Y.H., Kim J.K., Kim S.H.: A personalized recommender system based on web usage mining and decision tree induction. Expert Systems with Applications 23 (3), 2002, 329-342.
[20] ChoiceStream. Annual Choicestream Survey Finds Shoppers Avoiding Online Retailers that Deliver Poor Recommendations. December 2007, Cambridge, Massachusetts, USA.
[21] Cosley D., Lam S., Albert I., Konstan J., Riedl J.: Is seeing believing? How recommender interfaces affect users' opinions. CHI 2003, 2003, ACM, 585-592.
[22] Degemmis M., Lops P., Semeraro G.: A content-collaborative recommender that exploits WordNet-based user profiles for neighborhood formation. User Modeling and User-Adapted Interaction 17 (3), 2007, 217-255.
[23] Dolog P., Henze N., Nejdl W., Sintek M.: Personalization in distributed e-learning environments. The 13th Int. Conf. on World Wide Web, Alternate Track Papers & Posters, WWW 2004, ACM, 2004, 170-179.
[24] Dolog P., Kravcik M., Cristea A.I., Burgos D., De Bra P., Ceri S., Devedzic V., Houben G.-J., Libbrecht P., Matera M., Melis E., Nejdl W., Specht M., Stewart C., Smits D., Stash N., Tattersall C.: Specification, authoring and prototyping of personalised workplace learning solutions. International Journal of Learning Technology 3 (3), 2007, 286-308.
[25] Dolog P., Simon B., Nejdl W., Klobucar T.: Personalizing access to learning networks. ACM Transactions on Internet Technology 8 (2), 2008.
[26] Donaldson J.: A hybrid social-acoustic recommendation system for popular music. RecSys 2007, ACM, 2007, 187-190.
[27] Drachsler H., Hummel H., Koper, R.: Applying recommender systems to lifelong learning networks: requirements, suitable techniques and their evaluation. Journal of Digital Information (submitted), http://dspace.learningnetworks.org/bitstream/1820/1187/1/150108_Drachsler_JoDI.pdf.
[28] Farzan R., Brusilovsky P.: Social Navigation Support in a Course Recommendation System. AH 2006, LNCS 4018, Springer Verlag, 2006, 91-100.
[29] Felfernig A..: Knowledge-based Recommender Technologies Supporting the Interactive Selling of Financial Services. Chapter VI in Blecker T., Friedrich G.: Mass Customization Information Systems in Business, IGI Global, 2007, 122-135.
[30] Felfernig A., Gordea S., Jannach D., Teppan E., Zanker M.: A short Survey of Recommendation Technologies in Travel and Tourism., OEGAI Journal 25 (7), Oesterreichische Gesellschaft fuer Artificial Intelligence, 2007, 17-22.
[31] Felfernig A., Gula B., Teppan E.: KOBA4MS: Knowledge-based Recommender Technologies for Marketing and Sales. International Journal of Pattern Recognition and Artificial Intelligence, 21(2), 2007, 333-355.
[32] Géry M., Haddad M.H.: Evaluation of web usage mining approaches for user's next request prediction. WIDM 2003, ACM Press, 2003, 74-81.
[33] Goldberg K., Roeder T., Gupta D., Perkins C.: Eigentaste: a constant time collaborative filtering algorithm. Information Retrieval, 4(2), 2001, 133-151.
[34] Guo X., Lu J.: Intelligent e-government services with personalized recommendation techniques. International Journal of Intelligent Systems 22 (5), 2007, 401-417.
[35] Gupta A., Paul S., Jones Q., Borcea C.: Automatic identification of informal social groups and places for geo-social recommendations. International Journal of Mobile Network Design and Innovation 2 (3-4), 2007, 159-171.
[36] Harper F.M., Sen S., Frankowski D.: Supporting social recommendations with activity-balanced clustering. RecSys 2007, ACM, 2007, 165-168.
[37] Hsu M.-H.: A personalized English learning recommender system for ESL students. Expert Systems with Applications 34 (1), 2008, 683-688.
[38] Huang A.W., Sundaresan N.: Aurora: a conceptual model for Web-content adaptation to support the universal usability of Web-based services. CUU'00, ACM, 2000, 124-131.
[39] Hummel H.G.K., van den Berg B., Berlanga A.J., Drachsler H., Janssen J., Nadolski R., Koper R.: Combining social-based and information-based approaches for personalised recommendation on sequencing learning activities. International Journal of Learning Technology 3(2), 2007, 152-168.
[40] Iyer G., Soberman D., Villas-Boas J.M.: The Targeting of Advertising. Marketing Science, 24 (3), 2005, 461-476.

[41] Kamahara J., Nomura Y., Ueda K., Kandori K., Shimojo S., Miyahara H.: A TV News Recommendation System with Automatic Recomposition. AMCP '98, LNCS 1554, Springer Verlag, 1999, 221-235.
[42] Karamon J., Matsuo Y., Ishizuka M.: Generating Useful Network Features from Social Networks for Web 2.0 services. WWW2008.
[43] Kazienko P.: Filtering of Web Recommendation Lists Using Positive and Negative Usage Patterns. KES2007, LNAI 4694, Part III, Springer Verlag, 2007, 1016-1023.
[44] Kazienko P.: Product Recommendation in E-Commerce Using Direct and Indirect Confidence for Historical User Sessions. DS'04, LNAI 3245, Springer Verlag, 255-269.
[45] Kazienko P.: Usage-Based Positive and Negative Verification of User Interface Structure. ICAS 2008, 2008, IEEE, 2008, 1-6.
[46] Kazienko P., Adamski M.: AdROSA - Adaptive Personalization of Web Advertising. Information Sciences, 177 (11), 2007, 2269-2295.
[47] Kazienko P., Kiewra M.: Link Recommendation Method Based on Web Content and Usage Mining. IIS:IIPWM´03, Advances in Soft Computing, Springer Verlag 2003, 529-534.
[48] Kazienko P., Kiewra M.: Personalized Recommendation of Web Pages. Chapter 10 in: Nguyen T. (ed.) Intelligent Technologies for Inconsistent Knowledge Processing. Advanced Knowledge International, Adelaide, South Australia, 2004, 163-183.
[49] Kazienko P., Kołodziejski P.: Personalized Integration of Recommendation Methods for E-commerce. Int. Journal of Computer Science & Applications 3 (3), August 2006, 12-26.
[50] Kazienko P., Musiał K.: On Utilising Social Networks to Discover Representatives of Human Communities. Int. Journal of Intelligent Information and Database Systems 1(3/4), 2007, 293-310.
[51] Kazienko P., Musiał K.: Recommendation Framework for Online Social Networks. AWIC 2006, Studies in Computational Intelligence, Springer, Vol. 23, 2006, 111-120.
[52] Kazienko P., Musiał K., Juszczyszyn K.: Recommendation of Multimedia Objects based on Similarity of Ontologies. KES 2008, Springer Verlag, LNAI, 2008, in press.
[53] Kazienko P., Pilarczyk M.: Hyperlink Recommendation Based on Positive and Negative Association Rules. New Generation Computing 26 (3), May 2008, 227-244.
[54] Kim Y.S., Yum B.-J., Song J., Kim S.M.: Development of a recommender system based on navigational and behavioral patterns of customers in e-commerce sites. Expert Systems with Applications 28(2), 2005, 381-393.
[55] Kim K.-j., Ahn H.: A recommender system using GA K-means clustering in an online shopping market. Expert Systems with Applications 34 (2), 2008, 1200-1209.
[56] Kitts B., Freed D., Vrieze M.: Cross-sell: a fast promotion-tunable customer-item recommendation method based on conditionally independent probabilities. KDD, ACM, 2000, 437-446.
[57] Król D., Szymanski M., Trawiński B.: Web-Based Recommendation Strategy in a Cadastre Information System. AH 2006, LNCS 4018, Springer Verlag, 2006, 346-349.
[58] Krulwich B.: Lifestyle Finder: Intelligent User Profiling Using Large-Scale Demographic Data. AI Magazine, 18 (2), 1997, 37-45.
[59] Lai H., Yang T.C.: A Group-based Inference Approach to Customized Marketing on the Web - Integrating Clustering and Association Rules Techniques. System Science 2000, HICSS-33, IEEE Computer Society, 2000, p. 6054.
[60] Langheinrich M., Nakamura A., Abe N., Kamba T., Koseki Y.: Unintrusive Customization Techniques for Web Advertising. Computer Networks 31 (11-16), 1999, 1259-1272.
[61] Lavrenko V., Schmill M.D., Lawrie D., Ogilvie P., Jensen D., Allan J.: Language Models for Financial News Recommendation. CIKM'00, ACM 2000, 389-396.
[62] Lawrence R.D., Almasi G.S., Kotlyar V., Viveros M.S., Duri S.S.: Personalization of Supermarket Product Recommendations. Data Mining & Knowledge Discovery 5(1/2), 1001, 11-32.
[63] Lee H., Smeaton A.F., O'Connor N.E., Smyth B.: User evaluation of Físchlár-News: An automatic broadcast news delivery system. ACM Trans. on Information Systems 24 (2), 2006, 145-189.
[64] Lee J., Jun W.: An Adaptive School Web Site Construction Algorithm Using Association Rules. International Journal of Computer Science and Network Security, 8 (1), January 2008, 196-202.
[65] Lee T.-Q., Park Y., Park Y.-T.: A time-based approach to effective recommender systems using implicit feedback. Expert Systems with Applications 34 (4), 2008, 3055-3062.
[66] Lenar M., Sobecki J.: Using Recommendation to Improve Negotiations in Agent-based Systems. Journal of Universal Computer Science 13 (2), 2007, 267-285.
[67] Li J., Zaïane O.R.: Combining Usage, Content, and Structure Data to Improve Web Site Recommendation. EC-Web 2004, LNCS 3182, Springer Verlag, 2004, 305-315
[68] Li Q., Myaeng S.-H., Guan D.H., Kim B. M.: A Probabilistic Model for Music Recommendation Considering Audio Features. AIRS 2005, LNCS 3689, Springer, 2005, 72-83.

[69] Lin K.-L.; Hsu J.Y.-J.; Huang H.-S., Hsu C.-N.: A recommender for targeted advertisement of unsought products in e-commerce. CEC 2005, IEEE Computer Society, 101-108.
[70] Linden G., Smith B., York J.: Amazon.com Recommendations: Item-to-Item Collaborative Filtering. IEEE Internet Computing 7(1), 2003, 76-80.
[71] Linton F., Schaefer H.P.: Recommender Systems for Learning: Building User and Expert Models through Long-Term Observation of Application Use. User Modeling and User-Adapted Interaction 10 (2-3), 2000, 181-207.
[72] Liu D.-R., Shih Y.Y.: Integrating AHP and data mining for product recommendation based on customer lifetime value. Information & Management 42 (3), 2005, 387-400.
[73] McCarthy J.F.: The challenges of recommending digital selves in physical spaces. RecSys 2007, ACM, 2007, pp. 185-186.
[74] Millan M., Trujillo M.F., Ortiz E.: A Collaborative Recommender System Based on Asymmetric User Similarity. IDEAL 2007, LNCS 4881, Springer, 2007, 663-672.
[75] Moon S.K., Simpson T.W., Kumara S.R.T.: An Agent-based Customized Recommender System for Product and Service Family Design. The 2007 Industrial Engineering Research Conference, 2007, http://edog.mne.psu.edu/pdfs/IERC2007.44.w-Seungki.final.pdf.
[76] Montaner M., López B., de la Rosa J. L.: A Taxonomy of Recommender Agents on the Internet. Artificial Intelligence Review 19 (4), 2003, 285-330.
[77] Morzy M., Jezierski J.: Cluster-based analysis and recommendation of sellers in online auctions. TrustBus 2006, LNCS 4083 Springer, 2006, 172-181 .
[78] Nathanson T., Bitton E., Goldberg K.Y.: Eigentaste 5.0: constant-time adaptability in a recommender system using item clustering. RecSys 2007, ACM, 2007, 149-152.
[79] Ohsugi N., Monden A., Matsumoto K.: A Recommendation System for Software Function Discovery. APSEC 2002, IEEE Computer Society, 2002, 248-257.
[80] O'Sullivan D., Smyth B., Wilson D., McDonald K., Smeaton A.F.: Improving the quality of the personalized electronic program guide. User Modeling and User-Adapted Interaction 14 (1), 2004, 5-36.
[81] Pan S.-F., Lee J.-H.: eDAADe: An Adaptive Recommendation System for Comparison and Analysis of Architectural Precedents. AH 2006, LNCS 4018, Springer Verlag, 2006, 370-373.
[82] Pazzani M.: A Framework for Collaborative, Content-Based and Demographic Filtering. Artificial Intelligence Rev., 13 (5-6), 1999, 393-408.
[83] Perner P., Fiss G.: Intelligent E-marketing with Web Mining, Personalization, and User-Adapted Interfaces. Advances in Data Mining: Applications in E-Commerce, Medicine, and Knowledge Management, LNAI 2394, Springer Verlag, 2002, 37-52.
[84] Ponnada M.., Sharda N.: A High Level Model for Developing Intelligent Visual Travel Recommender Systems. Information and Communication Technologies in Tourism 2007, Springer, 2007, 33-42.
[85] Pronk V., Verhaegh W.F.J., Proidl A., Tiemann M.: Incorporating user control into recommender systems based on naive bayesian classification. RecSys 2007, ACM, 2007, 73-80.
[86] Schafer J.B., Konstan J.A., Riedl J.: E-Commerce Recommendation Applications. Data Mining and Knowledge Discovery 5 (1/2), 2001, 115-153.
[87] Semeraro G., Andersen V., Andersen H.H.K., de Gemmis M., Lops P.: User profiling and virtual agents: a case study on e-commerce services. Universal Access in the Information Society, 2008, http://www.springerlink.com/content/f8n7472646417u72/fulltext.pdf.
[88] Senecal S., Nante J.: The influence of online product recommendations on consumers' online choices. Journal of Retailing 80 (2), 2004, 159-169.
[89] van Setten M., Pokraev S., Koolwaaij J.: Context-Aware Recommendations in the Mobile Tourist Application COMPASS. AH 2004, LNCS 3137, Springer Verlag, 2004, 235-244.
[90] Sobecki J., Babiak E., Słanina M.: Application of Hybrid Recommendation in Web-Based Cooking Assistant. KES 2006, Part III, LNAI 4253, Springer, 2006, 797-804.
[91] Sobecki J., Szczepanski L.: Wiki-News Interface Agent Based on AIS Methods. KES-AMSTA 2007, LNAI 4496, Springer, 2007, 258-266.
[92] Spiliopoulou, M., Pohle, C.: Data Mining for Measuring and Improving the Success of Web Sites. Data Mining and Knowledge Discovery 5 (1/2) , 2001, 85-114.
[93] Srikant, R., Yang, Y.: Mining web logs to improve website organization. WWW 10, ACM Press, 2001, 430-437.
[94] Stewart A., Niederée C., Mehta B., Hemmje M., Neuhold E.: Extending Your Neighborhood - Relationship-based Recommendations for your Personal Web Context. ICADL 2004, LNCS 3334, Springer Verlag, 2004, 523-532.
[95] Symeonidis A. L., Chatzidimitriou K.C., Kehagias D., Mitkas P.A.: An Intelligent Recommendation Framework for ERP Systems. IASTED Int. Conf. on Artificial Intelligence and Applications, The 23rd Multi-Conference on Applied Informatics, IASTED/ACTA Press 2005, 715-720.

[96] Tang T.Y., McCalla G.: Smart Recommendation for an Evolving E-Learning System. AIED03, Online Supplementary Proceedings, http://www.cs.usyd.edu.au/~aied/vol10/vol10_TangMcCalla.pdf.
[97] Tan H., Lambrix P.: A Method for Recommending Ontology Alignment Strategies. ISWC 2007/ASWC 2007, LNCS 4825, Springer, 2007, 494-507.
[98] Thio N., Karunasekera S.: Automatic Measurement of a QoS Metric for Web Service Recommendation. ASWEC 2005, IEEE Computer Society, 2005, 202-211.
[99] Tran T.: Designing Recommender Systems for E-Commerce: An Integration Approach. ICEC'06, ACM International Conference Proceeding Series 156, ACM, 2006, 512-518.
[100] Umyarov A., Tuzhilin A.: Leveraging aggregate ratings for better recommendations. RecSys 2007, ACM, 2007, 161-164.
[101] Vellino A., Zeber D.: A Hybrid, Multi-dimensional Recommender for Journal Articles in a Scientific Digital Library. The 2007 IEEE/WIC/ACM International Conference on Web Intelligence and International Conference on Intelligent Agent Technology - Workshops, IEEE, 2007, 111-114.
[102] Wang Y.-F., Chuang Y.-L., Hsu M.H., Huan-Chao K.: A personalized recommender system for the cosmetic business. Expert Systems with Applications 26 (3), 2004, 427-434.
[103] Wang H.-C., Lee C.-S., Ho T.-H.: Combining subjective and objective QoS factors for personalized web service selection. Expert Systems with Applications 32 (2), 2007, 571-584.
[104] Winterboer A., Moore J.D.: Evaluating information presentation strategies for spoken recommendations. RecSys 2007, ACM, 2007, 157-160.
[105] Woerndl W., Eigner R.: Collaborative, Context-Aware Applications for Inter-networked Cars. WETICE 2007, IEEE Computer Society 2007, 180-185
[106] Wörndl W., Schüller C., Wojtech R.: A Hybrid Recommender System for Context-aware Recommendations of Mobile Applications. ICDE 2007, IEEE Computer Society, 2007, 871-878.
[107] Xu J.A., Araki K.: A Personalized Recommendation System for Electronic Program Guide. AI 2005, LNAI 3809, Springer Verlag, 2005, 1146-1149.
[108] Yager R.R.: Targeted E-commerce Marketing Using Fuzzy Intelligent Agents. IEEE Intelligent Systems 15 (6), 2000, 42-45.
[109] Yang W.-S., Dia J.-B.: Discovering cohesive subgroups from social networks for targeted advertising. Expert Systems with Applications 34 (3), 2008, 2029-2038.
[110] Yu Z., Lin N., Nakamura Y., Kajita S., Mase K.: Fuzzy Recommendation towards QoS-Aware Pervasive Learning. AINA 2007, IEEE Computer Society 2007, 604-610.
[111] Zaletelj J., Wages R., Bürger T., Stefan M., Grünvogel S.M.: Content Recommendation System in the Production of Multi-Channel TV Programs. AXMEDIS'07, IEEE Computer Society, 2007, 211-218.
[112] Zhang Y., Jiao J.R.: An Associative Classification-based Recommendation System for Personalization in B2C E-commerce Applications. Chapter V in Blecker T., Friedrich G.: Mass Customization Information Systems in Business, IGI Global, 2007, 107-121.
[113] Zhang D., Simoff S., Aciar S., Debenham J.: A multi agent recommender system that utilises consumer reviews in its recommendations. Int. J. of Intelligent Information and Database Systems, 2 (1), 2008, 69-81.
[114] Zhou B., Hui S.C., Chang K.: An intelligent recommender system using sequential Web access patterns. 2004 IEEE Conference on Cybernetics and Intelligent Systems, IEEE, 2004, 393- 398.

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-259

Code and Data Propagation on a PC's Multi-Agent System

Dariusz KRÓL and Aleksander LUPA
Wrocław University of Technology, Institute of Applied Informatics,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
e-mail: dariusz.krol@pwr.wroc.pl

Abstract. We present the concept of code and data propagation on a PC's multi-agent system. First, the idea of the prime numbers problem in distributed environment is defined and explained. Afterwards, the code and data propagation in a multi-agent system is described, which is in many cases based on the object migration. In addition, the experiments are conducted in a heterogeneous environment. The main aim is to find optimal configuration after minimal number of migrations for that platform.

Keywords. agent migration, propagation strategy, mobile agent

Introduction

In distributed computing it often happens that nodes in the network have different configuration and efficiency due to their CPU power, the amount of available memory, operating system used, and other factors. When nodes get their tasks, the distributing agent process usually does not know how much time it will take for each node to complete its task part. This information is often available after some time during the task solving. Moreover, during the execution it might happen that on one or more nodes there is a sudden computation power decrease, which might slow the entire system. This kind of environment requires an adaptive system that is able to dynamically load balance task solving [1].

A multi-agent technology is based on agents, which act as individuals [2], [3]. They have their own view of the environment and they can share it with other agents. This is also a technology that can be used to implement a distributed computing system. In the situation presented earlier, agents are able to react to a sudden environmental change. The question is how will they react and what is the intended result of their reaction?

Code and data propagation [4] is a technique that can be used in distributing task between computers in the network. This can happen before the experiment (like spreading object with a task to do over the network) and also during the experiment (using objects migration). If an object possesses information what has to be done in the task it can become a unit that will also possess information about result. When such objects are autonomous, they can perform a task simultaneously and then combine results. Of course, they can also migrate in order to complete the task on other computer. These three aspects combined give a description of a problem and its solution. There is a multi-agent system, where agents migrate within a heterogeneous

network. Agents could also be equipped with information about the task, which is distributed at the beginning between nodes in the network.

The concept of propagation is also common to distributed environments in the following forms: constraint propagation [5], [6], moving object [7], [8], computer virus [9], message dispatching [10], and many more.

From the definition agents in the multi-agent system are autonomous, but they also try to do what they were requested [11]. They are objects with some data, they possess a state, and they are able to perform operations. So they are built from code and data and they are the source of propagation.

In a multi-agent system there are three ways of code and data propagation [12]. An agent might send a message to another agent with a data in it (data propagation) and try to pursue the receiver to perform an operation, that a sender is performing (code propagation). The receiver, however, does not have to agree with this proposal. After sending a message an agent could be created to do the same job, as the message sender. Second way of code and data propagation relies on copying an agent. In this case the replica functions in the same way as the original one, possesses the same code and data. After the copy is made, agent starts its execution and functions in the same way as the original one. The third way consists in moving an agent between environments. An agent leaves one environment in order to resume its execution in another one.

1. Proposed Framework

The first idea was to find a problem that can be solved in a distributed way. The task had to have a possibility of being divided into subtasks, which are quite independent from each other and at the end, it is possible to merge all parts and get a full solution. Of course the main goal was not to find a very difficult task, because solving it is not the primary objective, but to find something that can represent a distributed task with many of its features. The main goal of a suitable task searching was to find one where code and data propagation could be introduced with a good probability of having benefits from doing so.

Code and data propagation (implemented by agent migration) should more or less function as a load balancing algorithm, levelling the amount of work on machines basing on their efficiency. Agent should represent a part of work, moving it from the slow computer to the faster one could balance the execution time.

1.1. Assumptions

There are several assumptions, which were made in order to focus on code and data propagation in a multi-agent system:

- The number of packets sent in the network equals the number of packets delivered. All packets are successfully delivered from source to the intended destination.
- There are no factors that can suddenly slow down the traffic in the network.
- There are no unexpected errors on the node computers that can interrupt systems normal functioning. On the other hand it is possible that a computer slows down as a result of another program running or a virus.

- There is no information at the beginning of a test regarding the configuration of nodes on which system is running. The only knowledge is that nodes are capable of solving the task.

The JADE multi-agent platform was chosen to be the environment for system implementation. One of the main reasons for choosing JADE environment is that it is used in many publications as a testing environment and it is considered as an efficient one. It is also widely accepted by scientists, which test their MAS project implementations on JADE. As one out many it is compliant with FIPA specifications and it is open source project. The implementation of this project is done in JADE v3.5.

1.2. Prime Numbers Problem Solving

The prime numbers problem solving was chosen for this project. The assumption is that at the beginning there is a range given into the system to calculate all prime numbers and write the results to the file. When the problem is being solved in the network all computers are calculating prime numbers and as soon as possible they are sending the solution to the one, which is responsible for saving them to the file. This is done to have the packets in the network transferred all the time. Having it solved this way makes the solution efficient, because while examining candidates for prime number, the network connection is also used.

This problem was chosen to be solved in a distributed way also because it is computation-focused task. Agent here is given a range of numbers to examine and from the result point of view, it does not matter, on machine it is, and here comes the load balancing part. There is a possibility to move agents from the slow computers to the faster ones, using though heterogeneity for faster task solving completion.

The algorithm presented below is very easy. Of course, there are more efficient ways to calculate prime numbers, but the goal is not to calculate them as fast as possible, but to simulate solving the task, which can be distributed.

For each x from given range [a,b]
  Begin
    Calculate square root c = sqrt(x);
    Check if any number from range [2,c] divides x without reminder
    If no then add x to the prime numbers list
  End

This approach has several features. One of them is that the range to search the primes from can be divided, and the results lists can be merged at the end giving the full solution to the problem. In order not to create a bottle-neck, partial results are being sent continuously to be saved to the file.

1.3. The Network

The network is composed of two kinds of computers: nodes and a broker. A node is a computer, on which agents reside, which is supposed to calculate prime numbers. A broker is a computer, which distributes the task into the network and saves the results into the file. It is not destined to calculate prime numbers.

The algorithm for distributing the task over the network is quite natural. When the broker gets the task order it knows how many nodes are available, because when they enter the network, they send a ready message to the broker. Because at the beginning there is no information about the network, primes range search is divided equally by

the number of nodes and sent to them as a task request. While calculating, nodes send found primes to the broker by portions. This makes the data flow fluent through the time of the task solving.

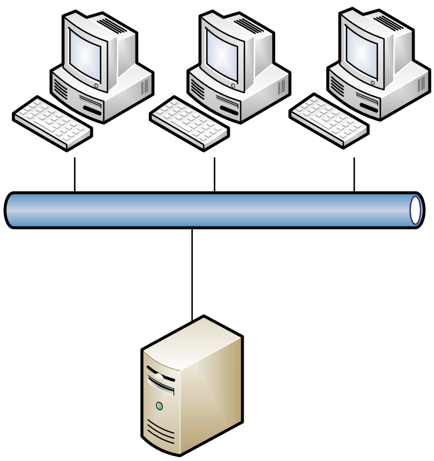


Figure 1. Network topology

From the JADE perspective of agent containers there are two main types of containers. One of them is the "Main-Container" - an equivalent of a broker in the described network. There is only one instance of this container. On the other hand, there are simple "Containers", with agents calculating prime numbers. There can be many instances of "Containers" and, moreover, there can be many of them on one computer. In the network they are equivalent to a node with agents.

2. Agent Migration

2.1. Agents in the System

There are 5 agents in the implemented system: Broker, Saver, Coordinator, Primers and Migration Coordinator. Each of them has a specific characteristics and it cooperates with some other agents to achieve a goal.

Broker is the core agent of the system and there is only one in the entire network. It is placed in the "Main-Container". It has two main tasks:

- Task distribution. After getting a message with the range of number to examine for primes it divides it equally and sends parts to Coordinators.
- Experiment completing. Agent waits for all Coordinators to get end of work messages from them. Then, it ends the experiment.

Saver is an agent that is connected to the file. There is only one agent of that kind and it resides in "Main-Container". Its main responsibility is receiving messages with results from Primers and saving them to the file.

Primer is the key agent in the system. It searches prime numbers from the given range of numbers and works in parallel with other Primers. It is often called as the

agent in this paper, because it is the main one, which actually does the job of calculation. As many other agents it sets its own parameters: migration type and primes range search unit. It has several assignments:

- Calculating primes. Agent searches through the entire range of number in order to examine every number if it is a prime number. It adds them to the list, which is sent to the Saver agent to write these results to the file. Agent does it in portions. It means it searches primes within a range length set, which is a parameter and then, sends results and checks if it has a message incoming from other agents.
- Progress report sending. Agent sends a report to the local Coordinator with amount of number checked for prime property.
- Migration. Agent migrates to another environment (container) and continues its execution there.
- End of work communication. When the range of numbers has been examined and there are no more numbers to check, agent sends a message to the local Coordinator.

Coordinator is responsible for managing Primers in the container. It is strictly an agent responsible for coordination and info exchange for a certain node. There is always one Coordinator in each container (except Main-Container). It has several assignments:

- Local task distribution. After getting a range of numbers to examine, agent divides it equally into parts and sends them to Primers available at current location.
- Time till end estimation. Agents estimates time till the end of work on a local node based on the speed of checking number and the amount of them for examination for prime property.
- Local migration coordination. Agent coordinates Primers migration. It counts how many agents will arrive, how many will leave and also chooses the ones that will leave and sends message to them. If a Coordinator checks that some agents cannot leave and it sends reject message to the Migration Coordinator about this agent and sends a correction message to the agent's destination.
- Gathering state information. After the migration phase is over agent sends progress report request to Primers in order to have a basis for the next time till the end of estimation.
- Job completion communication. When all Primers finish their job agents send a message to Migration Coordinator and Broker.

Migration Coordinator is responsible for coordinating the migration in the system. It manages reports and migration proposals sent to Coordinators. It also calculates how the migration in the system should look like. The main objective is to control the steps regarding the migration process during the experiment. There is only one Migration Coordinator in the network and it resides in the "Main-Container". In the experiment agent has several assignments:

- Migration calculation. Agent calculates the new state for the network (number of agents in each Container).
- Migration proposals sending. Agent sends proposals of migration (giving the destination for each migrating agent) to Coordinators.
- Gathering reports from nodes. Agent gathers reports with estimated time and active agents from all nodes and gathers this information in a structured way.

- Report requests sending. Agents sends report request to Coordinators.
- Gathering migration finishing reports. Agent gathers migration completion reports from all coordinators. This is one of the tasks, which guards the phase order in the experiment.

2.2. Important Parameters

There are several important parameters in the system, which are crucial in the experiment, because they affect both the efficiency and execution process:

- Number of agents on one node. This is the number of Primers that will be initiated in each Container. This number multiplied by the number of Containers gives the overall amount of Primers in the system.
- Primes range search unit. This is one of the most important parameters in the system (also called elementary primes range search). When a Primer is searching for prime numbers, it has to have a range of numbers to search from. This parameter is a number of numbers, which are checked at one time. After a Primer checks a unit of numbers, it generally checks if there are messages waiting to be handled. This parameter is also responsible for how often the save messages (a message sent from Primers to Saver agent) are generated. Every Primer agent, after checking a range unit, sends a message to the Saver agent in order for saving the results (prime numbers). After primes are sent the list of prime numbers held by the Primer agent is cleared.
- Report time. This parameter is given in a number of seconds, which have to elapse from the end of migration phase (when all Coordinators report local migration has finished) till the next report request sending by Migration Coordinator.
- Node threshold. This parameter is given in a number of seconds. It is strictly connected with migration process. Coordinators are sending reports to Migration Coordinator - they send the estimated time till the job ends on the node and number of agents. When migration calculation process starts, it needs all these values, but it does not include a node into calculation when its estimated time is lower than node threshold. Then this node is also omitted when calculating average active node time, sum of agents and sum of proposed agents. This parameter is given in order to avoid unnecessary migration. This migration is when a node has already finished its job and afterwards there are new agents arriving with some work to do. This parameter could also be used to anathematize nodes from migration earlier.
- Valid time. This parameter is responsible for minimal time trial when estimating time till the end of experiment.

2.3. Migration Calculation

This algorithm is in other words also the process of assessing the ability of a node to perform a task. The key point here is that the estimation is based on the previous efficiency of a node when performing a task. It is compliant to the assumption that at the beginning of an experiment the ability of nodes to perform the task is unknown.

When a Coordinator gets all progress reports it estimates time till the end of experiment based on the data it has. This is a very important moment in the whole

algorithm. Time till end is estimated based on a sample from the previous change in the environment for local node – so from the last migration or the beginning of the experiment. Basically, when Primers are reporting, they deliver two parameters - how big is the range of numbers they have to examine and how many numbers have already been checked for prime property. All Primers report that there are two sums calculated: a sum of numbers to check and a sum of numbers that already have been examined. Based on the new and old report there is a quantity of numbers calculated that have been examined. Then, the time from the last change in the environment is calculated. Based on these two values the speed is calculated:

$$speed = \frac{currently\ checked\ numbers - previously\ checked\ numbers}{current\ time - last\ change\ time}, \tag{1}$$

$$estimated\ time = \frac{quantity\ of\ numbers\ to\ ckeck}{speed}. \tag{2}$$

The first formula is changed only in one case. If there have been already some migrations, but the last saved change was before the "valid time", then the first formula is calculated from the beginning of an experiment. Valid time is one of the parameters for checking while doing the time estimation. If the last change was later then the valid time margin (counting from now during the experiment) then estimated time is set from the beginning of an experiment. Time margin is calculated as the current time minus the valid time. For example, if the valid time was set to 1 second, then if the last change was conducted later then time estimation is calculated from the beginning of an experiment.

After the Migration Coordinator gets all reports form Coordinators it starts the algorithm. During that time all Coordinators wait for messages - they go to the ready state. When the agent network is being created at the beginning of an experiment, Migration Coordinator agent creates a list of objects describing nodes. This list is used for migration calculation but also for remembering names, locations in the network and for information if a node has to report back after migration finishing. As the reports arrive the information in the list is being refreshed.

For each node we have data on the estimated time and the list is sorted beginning with the shortest time till the end of experiment. Then for each node there are some values calculated:

$$agent\ value = \frac{node\ time}{number\ of\ node\ agents}, \tag{3}$$

$$agent\ change = \frac{average\ active\ time - node\ time}{agent\ value}, \tag{4}$$

$$proposed\ agents = node\ agents + agent\ change. \tag{5}$$

The first value is a measure of how much time from the estimated till the end falls to one agent. The second value is a bit more complicated. In this algorithm there is such a value as average active time. Active here means for those nodes only, where migration can take place or in other words, which have the estimated time higher than the node threshold parameter. So the goal is here to calculate how many agents should be on this node to have the time as close to average as possible. The assumption here is that an agent (Primer) represents some work to do and if there was a certain number of agents, then the time would be close to average. Having this simple assumption, the

number of proposed agents for each node is calculated. If for example a node has a time lower than average – then there should be more agents and the agent value change is greater than zero. If not then some agents should migrate from this node. But after calculating the proposed agent number there is a possibility that there should be more or less agents than currently is, so it is necessary to correct this number on each node by sum of agents divided by sum of proposed agents:

$$proposed\ agents = proposed\ agents * \frac{current\ agents\ sum}{sum\ of\ \ proposed\ agents} \qquad (6)$$

After this process agents are distributed according the resources (agents) available in the system. But there is a possibility that still the sums of agents and proposed agents are not equal, so there is correction algorithm, that makes these sums equal by adding or subtracting proposed agents for each nodes starting from those that have the biggest number of proposed agents.

After executing this algorithm a list of proposed migrations is created. Building this list is based on making equal the number of agents and proposed agents on a node (in the node information list) possessed by Migration Coordinator. Agents always migrate from the node that has the biggest estimated time to the node that has the lowest estimated time. Migration proposals' sending is only sending the messages that mirror the list of proposed migrations along with setting the obligation to report when migration is finished for those nodes that are accepting agents and for those nodes from where agents will depart.

3. Performance Evaluation

The main purpose of the experiments is to show, that code and data migration helps to improve efficiency for task completion when a network is composed of computers, which are not homogenous – they have different configuration or/and different computing power at the moment. In order to do this there has been a set of experiments conducted.

The rule in tests is to propagate the optimal values of one parameter to the next test and in that way find the optimal configuration. Of course many parameters can depend on each other, so the sequence of experiments is based on observations.

The test-bed is a home network, which is composed of four computers with different configuration each:

- Computer 1: laptop; Intel Core Duo 1.86 GHz (T2350), 2 GB RAM, Windows Vista Business SP1,
- Computer 2: laptop; Intel 1.86 GHz, 448 MB RAM, Windows XP Professional SP2,
- Computer 3: AMD Athlon XP 3200+ 2.2 GHz, 1 GB RAM, Windows XP Professional SP2,
- Computer 4: AMD Athlon XP 1700+ 1.44 GHz, 256 RAM, Windows XP Professional SP2.

The network is Ethernet 100Mb/s. All computers are connected to the Router, which is often used for home networking and small business networking. During the experiment, there were only few problems with the network, but generally it is a stable

environment. Later this environment will be called home environment, when describing experiments and their location.

3.1. Computer Efficiency Investigation

On each computer the task was completed, without connecting them into a network. Calculating and saving was conducted on the same machine. Only one JADE container (Main-Container) was used in order to exclude the impact of two JADE containers held on one machine. The range of numbers to search was the same for each computer. Experiment parameters are as follows: 20 agents on each node, primes search range unit of 1000. The range of numbers to check for primes was from 1 billion to 1.003 billion.

Goals:

- Investigating computers' solo efficiency task solving.
- Discovering the connection between agent number, processor type and execution time.
- The initial comparison of computers' efficiency.

Table 1. Execution times for all computer configurations

Agents / Computer	1	2	5	10	20	50	100	200	500	1000
Computer 3	326.9	328.7	328.1	325.1	321.9	318.8	318.4	318.6	324.9	329.6
Computer 1	181.9	92.3	93.9	93.9	93.6	94.1	93.9	93.8	95.1	95.6
Computer 2	184.1	185	185.8	184.9	185.1	185.3	187.2	188.5	193.2	197.5
Computer 4	479.1	478.9	481.1	481.1	481.4	482.1	482.5	483	493.3	511.6

The table above presents the execution time for all computer configurations. The number of agents has been set to get lower efficiency at the end of graph. Also there is one effect that was unexpected before running the experiment. For each computer there is a certain pattern which is also shown on the graph below: The execution time for one agent primes numbers calculation is lower than in situation when there are few of them. Afterwards the situation is different for different computers. For the following chart the execution time for one Primer is the global minimum. For example for the strongest machine the global minimum is for 2 agents. But the main trend is that for several agents the execution time rises and then it falls down (Computer 3) or keeps steady (Computer 4). When number of agents is becoming much higher the execution time is getting bigger.

The results in the table above are the average of 3 experiments on each computer. It is very important to mention, that they were very accurate – especially when the agent number was low. For the strongest computer the deviation was 0.1 s and for the weakest computer it was 0.5 s at maximum. It proves that the environment was very stable and from the project architecture we know that it was static. That is the reason of so close results in terms of each other.

It is very important to mention that deviations between 1 or 2 agents (except computer with core duo processor) and for example 50 agent in not big. The maximum difference value is 1%. It is not much and in many experiments, it could be perceived as a measurement error, but we know in this case it is not.

The comparison between computer efficiency delivers some knowledge about their "power" for task solving. In home environment, the difference between computers is

very big. The strongest machine is circa 5 times better than the weakest one, so the variety is big, which is well visible in the next experiments with migration.

The comparison between computers of the same architecture goes accordingly to the CPU configuration, but when it comes to the notebook and desktop computer comparison, it is not equivalent. The CPU configuration moves to the execution time: twice better CPU configuration – twice faster as experiment time. The problem is the comparison between the strongest desktop computer and the second notebook. Technically the first one should be better, because it has both stronger CPU and more RAM memory, but it is almost twice slower in task solving. Comparing home desktop computers, they seem to match each other in connection of CPU power (2.2 / 1.44 ~ 1.52) and execution time (for one Primer: 479.1 / 326.9 ~ 1.47). Comparing home strongest desktop computer and university's desktop computer, they seem to match also. CPU power (2.81 / 2.2 ~ 1.28) and execution time (for one Primer: (326.9 / 256.7 ~ 1.27).

There are several conclusions that could be drawn from the experiment:

- Execution time is strongly connected with the CPU configuration and it seems to be independent from the RAM memory capacity.
- The execution time is approximately a constant function with the connection to the number of Primers on a computer. This is the most important conclusion from this test and it will be used many times in the following experiments.
- The Intel Core Duo processor functions best when there are at least two agents calculating prime numbers. Computers with single core do not exhibit such a phenomenon.
- There is a phenomenon connected with number of agents and execution time and after the review of the result table the optimal agent number is about 20. So this amount of agents will be used for the following experiments.

3.2. Maximum and Minimum Strategies

The experiment is conducted in order to investigate the minimum and maximum strategy of distributing the task. The maximum strategy is here defined as giving the same task at first to the strongest computer, then the second computer, third and fourth. The strongest here means with the best efficiency. The minimum strategy works exactly in the opposite way – from the weakest computers to the strongest. Because there are no other computers, one of them has to perform file operations. In the minimum strategy, it is the weakest one, and in the maximum strategy, it is the most efficient one. The order here is set as follows: Core Duo 1.86 GHz, 2 GB; 1.86 GHz, 448 MB; 2.2 GHz, 1 GB; 1.44 GHz, 256 MB (also called Computer 1, 2, 3 and 4). After giving the same task to all computers with different order, migration is introduced in order to gain some profit in execution time.

Experiment parameters are as follows: 20 agents on each node, primes search range unit of 150, report time set as 5 seconds, node threshold set for 15 seconds, the default migration type is move. The range of numbers to check for primes was from 1 billion to 1.003 billion.

Goals:

- investigate execution time in connection with maximum and minimum strategy,

- investigate the time of computers reporting end of their part in connection with the execution time,
- investigation of lowering execution time by introducing migration,
- compare the time of computers reporting in connection with and without migration.

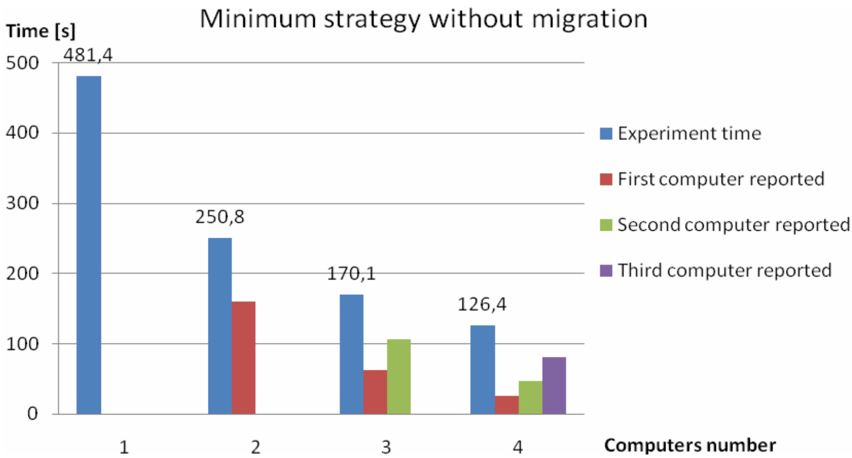


Figure 3. Minimum strategy without migration

Figure 3 shows the decreasing execution time along with introducing computers into the network by minimum strategy approach. The execution time here is increased by close to the optimal function so the number of execution time of one computer divided by the number of computers. The difference between the result and the estimated time is caused here by the file operations performed by the weakest computer. The time of computers reporting before the end of experiment is connected with their efficiency tested in the first experiment. The key point here is that there are big differences between first computer reporting ready state and execution time. The conclusion is that this environment needs introducing the migration.

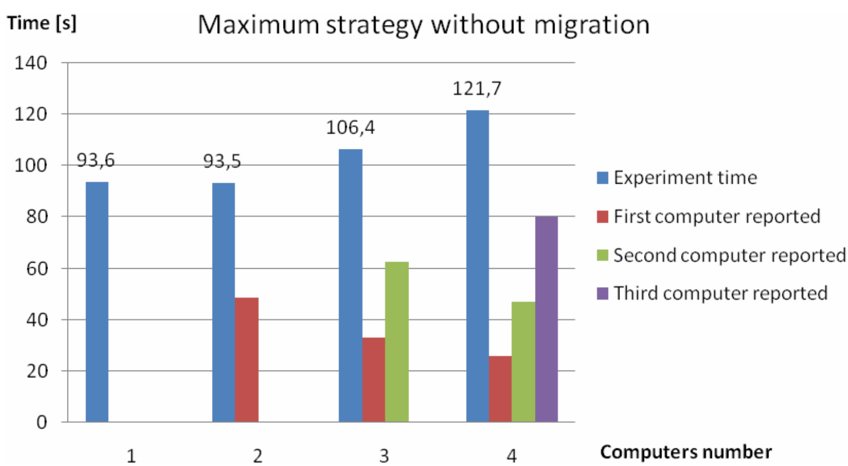


Figure 4. Maximum strategy without migration

Figure 4 shows the increasing execution time with introducing computers to the network. In the strategy we can see that the strongest computer is able to do the task faster than all 4 computers combined. It is because its efficiency is much higher than that of the others. With introduction of the first computer, it is twice slower than the strongest one. There is no profit. It is reasonable, because the range of candidates to search id divided by two, so the twice weaker computer will do it in the same time as twice stronger computer with twice longer set of numbers. With introduction of third and fourth computer, the execution time increases, because of the differences in computers' efficiency. The difference in execution time for four computers between two strategies is most probably caused by the file operation done by different computers. For the one with 1.44 GHz, 256 MB RAM configuration it is more time consuming than for Core Duo 1.86 GHz, 2 GB RAM.

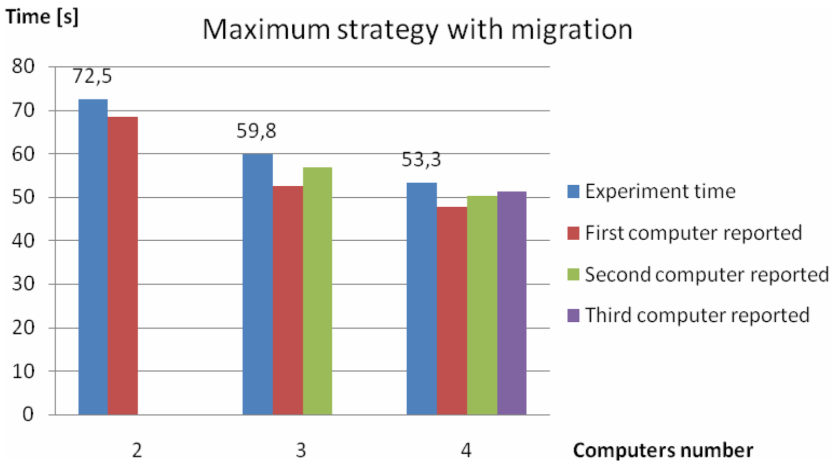


Figure 5. Maximum strategy with migration

Figure 5 shows the decreasing execution time with introducing computers into the network. The experiment time decreases slowly but in the end the profit is from 121 s without migration to 53 s with migration. The profit is around 56%, which is a very good result. Of course, when introducing migration there were some difficulties, because the execution time is rather short and agent migration takes time. For both strategies, multiple agent migration was used with 5 second delay between the end of reporting and the next report sending. This time, comparing to the overall execution time, is big, because some amount of work is already done before migration starts. Still the differences between experiment time and first computer reported are below 20%, which might be considered good because of short experiment time.

For minimum strategy the execution time decreases in a very fast way. Experiment time decreases almost twice with adding a computer into the network and, like in the previous strategy, the first computer reported time is close to the last one. The difference between first computer reported and the last one is similar to the maximum strategy with migration.

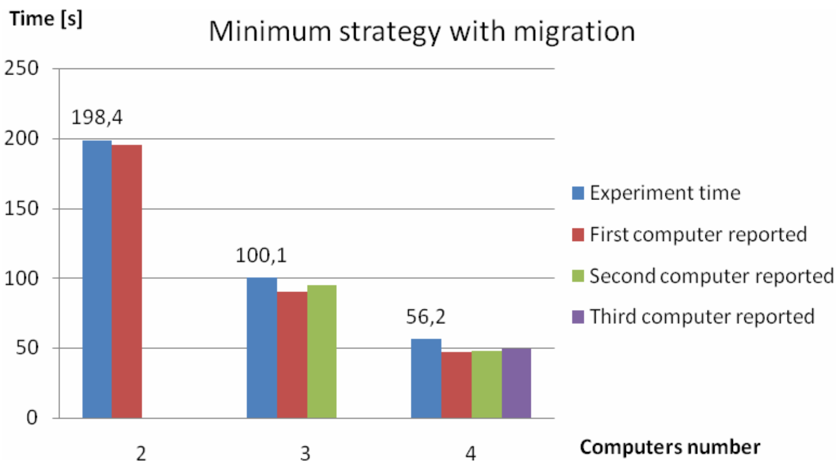


Figure 6. Minimum strategy with migration

There are several conclusions from maximum and minimum strategy testing:

- In the experiment without migration the execution time is determined by the slowest computer in the network.
- The sequence of computers reporting in experiment without migration is strictly connected with their efficiency. In home network the strongest computer does the work faster than four machines.
- In the heterogeneous network agent migration brings profit.
- The longer execution time, the greater profit from migration in a heterogeneous network.
- In heterogeneous network file saving operation should be done on the strongest computer.

3.3. Multi-Agent Migrations

This experiment is the next from the optimum searching with both home and university environment. This time multi-agent migration is used without any limits. The default type of migration is move. Again, on each node, there are 20 agents, end time equals 20 seconds and node threshold is set for 15 seconds. In this experiment there are two main parameters investigated: the number of agents on a node and report time. It is important to mention, that without the migration the results is 1022 s.

Goals are the following:

- Investigate the report time parameter and find the optimal value for multi-agent migration.
- Investigate the relationship between report time, migrations and report messages.
- Find the optimal number of agents on a node.
- Investigate the relationship between agents and migrations in the experiment.
- Investigate the relationship between migration type and execution time.

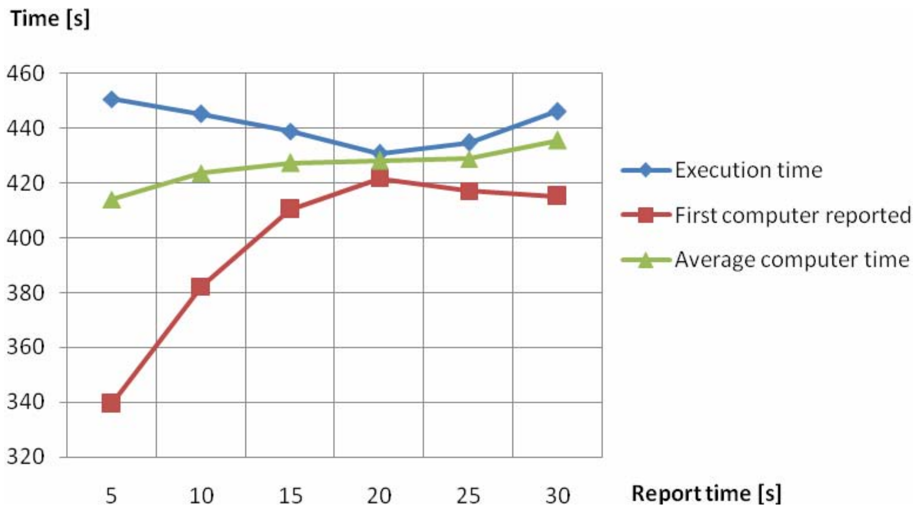


Figure 7. Relationship between computer time and report time

From the chart above, we can see that execution time and first computer reported meet in the minimum for 20 seconds of delay report. Execution time between all results does not differ so much (the difference between maximal and minimal experiment time is about 20 seconds, which is less that 5% of the whole experiment time). The first reported computer values have much bigger amplitude, but in this function, we are looking for local maximum, because when all computers finish in a comparable time, then, we are sure that migration was successful.

The explanation of these results is quite logical. When the report time is low, than the environment is more dynamic, the samples for time estimation are less reliable and so migration is not so accurate. The result of inefficient migration is that some computers may report very early during the experiment. When the report time is too high then the time is measured more precisely, but there is not enough time for reactions at the end of experiment, when some computers are finishing early and so there might me a difference between first reported and execution time. Another reason is very big report time might not increase the estimation accuracy so much.

The relationship between report time and migrations is definitely not regular (Figure 8). Migrations seem to find their optimum around 50 of them. The most part of this number is probably set at the beginning of an experiment and during it there are some minimal corrections – little agent movement. If the environment is more dynamic than there is a big movement at the beginning and there are still some migrations during the experiment. And so, with more accurate time estimations the migration number decreases.

Figure 9 shows how the number of agents influences the execution time, the time of first computer reporting, and the average of all four computers. The three measured values form a special formation, which is similar to a channel. The blue line is the forming the top of it, the red line the bottom and between them there is a green line, which is almost in the middle between them.

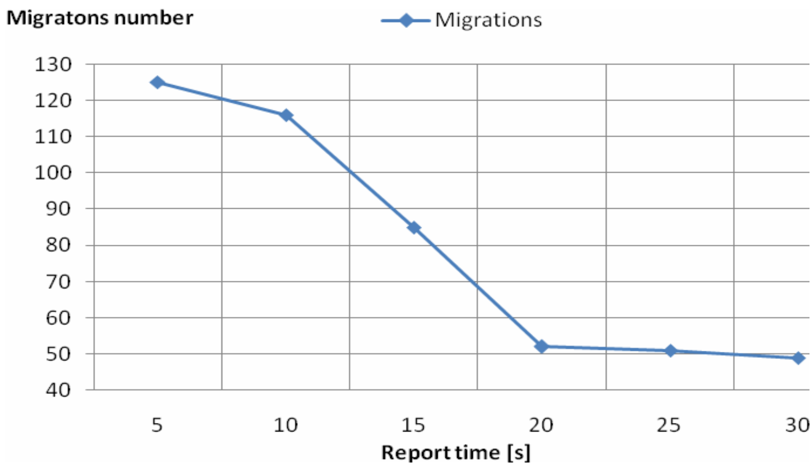


Figure 8. Relationship between report time and migrations number

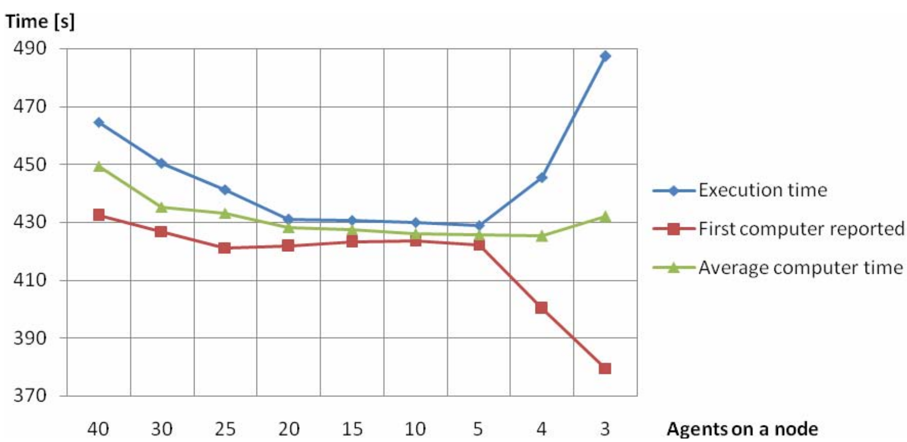


Figure 9. Relationship between number of agents on a node and computer time

On the left side of the chart (from 40 agents on a node), there is a descending trend of all three values, which ends for a number of agents equal to 20. This channel is not only descending but also the space range is descending. It ends for the value of 20, because then all values more or less enter the horizontal channel, which also forms the minimum of all values in the graph. All three lines are near each other – their results are comparable. For four agents on a node the channel opens and the lines disallow from each other.

The chart shows that the minimum, the best number of agents on a node is a set – from 5 to 20, which is quite a big range. Both these numbers have their advantages. If there are not many agents, there are less migrations and computers focuses more on calculating the prime numbers, but there are not enough agents to mirror the computer power in a precise way. When there are more agents on a node, mirroring the efficiency is easier, but there are more migrations and computer focus more on migration process.

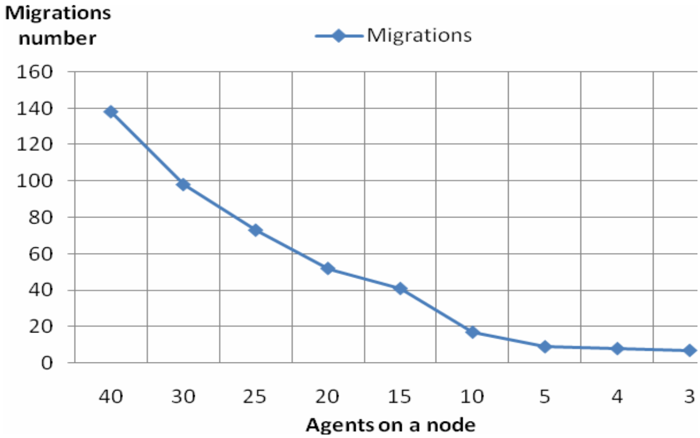


Figure 10. Relationship between agents on a node and migration number

Figure 10 shows that migration number is strongly correlated to the report messages. Both lines have similar shape and also have the same special points (for 15 agents on a node). Because the environment is dynamic, what cannot be easily explained, it should be treated as a consequence of dynamic interactions between agents and time estimation.

Analysing the chart, one can conclude that the migration number is more correlated to the agent number on a node, than to the report messages. The irregular formation in a line is less visible. If the point for 15 agents on a node was a little bit lower, the line would be smooth. Of course, this point could also be a consequence of dynamic interaction in the system. On the other side, the pattern of report messages is constructed of two parts (to the right from 15, and to the left of 15 agents on a node). Of course, it could also be a consequence of dynamic interactions, but it is less probable, that it happened for so many tests. Generally, the fewer agents are on a node, the fewer migrations.

There are several conclusions from this experiment:

- The longer the report time, the more accurate time is to estimate to the end of experiment for each node.
- The shorter the report time, the more dynamic the environment is.
- The optimal report time is around 20 seconds and above this number the environment also behaves in more stable way.
- Migration type seems to have little influence on the execution time for multi-agent migration.
- The optimal number of agents on a node is between 5 and 20.
- The fewer agents in the system, the fewer migrations.

4. Conclusions

The experimental section of this work has proven that code and data propagation in a multi-agent system (implemented by agent migration) increases the efficiency in task execution. The decrease in execution time in home environment was 58% (from 1022 s to 429 s). It would not be possible if the hypothesis about the number of agents

working on a node was proved: the number of agents on a node does not affect its efficiency when solving this task type (it was proved in computer efficiency experiment for low numbers of agents – up to 100).

Other important conclusions are as follows:

- The execution time is mainly dependent on CPU configuration. Desktop computer and notebook architecture differences result in different efficiency (the relation between CPU configuration does not match the efficiency in task solving).
- The more agents in the system, the more dynamic the environment is and the more migrations take place.
- The number of data transferred through the network was too small to have an impact on the execution time.
- The optimal configuration is when all computers finish their tasks in similar time.
- The closer the migration phases are in time, there are more migrations in the system. To make the system more stable migration phases have to be distant in time. It also affects the number of messages in the system – the more stable the system is, the lower communication costs are.
- The lower number of migrations, the shorter the execution time is (assuming there is enough agents in the system, that are able to cover differences in computer efficiency). The optimal number of agents seems to be between 5 and 20.

References

[1] S.S. Fatima, M. Wooldridge: Adaptive task resources allocation in multi-agent systems, *Proceedings of the Fifth International Conference on Autonomous Agents* (2001), 537-544.

[2] M. Kolp, P. Giorgini, J. Mylopoulos: *Multi-Agent Architectures as Organizational Structures, Autonomous Agents and Multi-Agent Systems* 13 (2006), 3-25.

[3] M. Wooldridge: *An Introduction to MultiAgent Systems*, Chichester, John Wiley & Sons, 2002.

[4] D. Krol: A Propagation Strategy Implemented in Communicative Environment, *Lecture Notes in Artificial Intelligence 3682* (2005), 527-533.

[5] S. Benbernou, M.S. Hacid: Resolution and Constraint Propagation for Semantic Web Services Discovery, *Distributed and Parallel Databases* 18 (2005), 65-81.

[6] M. Delafosse, A. Clerentin, L. Delahoche, E. Brassart and B. Marhic, Multi sensor fusion and constraint propagation to localize a mobile robot, *IEEE International Conference on Industrial Technology* (2004), 66-71.

[7] H. Güting, T. de Almeida, Z. Ding: Modelling and querying moving objects in networks, *The VLDB Journal* 15 (2006), 165-190.

[8] N. Krivokapic, M. Islinger and A. Kemper: Migrating Autonomous Objects in a WAN Environment, *Journal of Intelligent Information Systems* 15 (2000), 221-251.

[9] G. Serazzi, S. Zanero: Computer Virus Propagation Models, *Lecture Notes in Computer Science* 2965 (2001), 1-25.

[10] D. Krol, M. Zelmozer M., Structural Performance Evaluation of Multi-Agent Systems, *Journal of Universal Computer Science* 14 (2008), 1154-1178.

[11] N. Jennings, K. Sycara and M. Wooldridge: A roadmap of agent research and development, *Autonomous Agents and Multi-Agent Systems 1* (1998), 7-38.

[12] F. Bellifemine, G. Caire and D. Greenwood: *Developing Multi-Agent Systems with JADE*, Liverpool University, UK, 2001.

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-276

Cluster-Centric Approach to News Event Extraction

Jakub PISKORSKI [a,1], Hristo TANEV [a], Martin ATKINSON [a], Erik VAN DER GOOT [a]

[a] *Joint Research Centre of the European Commission*
Institute for the Protection and Security of the Citizen
Via Fermi 2749, 21027 Ispra, Italy

Abstract. This paper presents a real-time and multilingual news event extraction system developed at the Joint Research Centre of the European Commission. It is capable of accurately and efficiently extracting violent and natural disaster events from online news. In particular, a linguistically relatively lightweight approach is deployed, in which clustered news are heavily exploited at all stages of processing. The paper focuses on the system's architecture, real-time news clustering, geo-locating clusters, event extraction grammar development, adapting the system to the processing of new languages, cluster-level information fusion, visual event tracking and accuracy evaluation.

Keywords. Event extraction, shallow text processing, information aggregation

Introduction

Nowadays, massive amount of information is transmitted via Web, mostly in the form of free text. Recently, we have witnessed an ever-growing trend of utilizing natural language processing (NLP) technologies, which go beyond the simple keyword look-up, for automatic knowledge discovery from vast quantities of textual data available on the Web.

This paper reports on the multilingual event-extraction system developed at the Joint Research Centre of the European Commission for extracting violent and natural disaster event information from on-line news articles collected through the Internet with the Europe Media Monitor (EMM) [1], a web based news aggregation system, which regularly checks for updates of news articles across multiple sites in different languages. Gathering information about crisis events is an important task for better understanding conflicts and for developing global monitoring systems for automatic detection of precursors for threats in the fields of conflict and health.

Formally, the task of event extraction is to automatically identify events in free text and to derive detailed information about them, ideally identifying *Who did what to whom, when, with what methods (instruments), where and why*. Automatically extracting events is a higher-level information extraction (IE) task which is not trivial due to the complexity of natural language and due to the fact that in news a full event description is usu-

[1] Corresponding Author: Jakub Piskorski, E-mail: jakub@piskorski@jrc.it

ally scattered over several sentences and articles. In particular, event extraction relies on identifying named entities and relations holding among them. The research on automatic event extraction was pushed forward by the DARPA-initiated Message Understanding Conferences[2] and by the ACE (Automatic Content Extraction)[3] program. Although, a considerable amount of work on automatic extraction of events has been reported, it still appears to be a lesser studied area in comparison to the somewhat easier tasks of named-entity and relation extraction. Precision/recall figures oscillating around 60% are considered to be a good result. Two comprehensive examples of the current functionality and capabilities of event extraction technology dealing with identification of disease outbreaks and conflict incidents are given in [2] and [3] respectively. The most recent trends and developments in this area are reported in [4]

Due to our requirement that the event extraction system must be multilingual and easily extandable to new domains a linguistically relatively lightweight approach has been chosen. In particular, we take advantage of clustered news data at several stages of the event extraction process. As a consequence of this cluster-centric approach, only a tiny fraction of each news article is analyzed. Further, a cascade of finite-state extraction grammars is deployed in order to identify event-related information. These patterns are semi-automatically acquired in a bootstrapping manner, again via utilization of clustered news data. Exploiting clustered news intuitively guarantees better precision. Since information about events is scattered over different articles single pieces of information are validated and aggregated at cluster-level. One of the main prerequisites for the ability to digest massive data amounts in real time is efficient processing. Therefore, an in-house extraction pattern matching engine has been developed in order to find a good trade-off between terse linguistic descriptions and efficient processing. The system is fully operational since 2007 and supports event extraction for several languages. The results are viewed in real time via a publicly accessible web page. An empirical evaluation revealed acceptable accuracy and a strong application potential. Although our domain centers around the security domain, the techniques deployed in our system can be applied to other domains, e.g., tracking business-related events for risk assessment.

The rest of this paper is organized as follows. First, in section 1 the architecture of our live event extraction processing chain is described. Next, section 2 gives an overview of the real-time news clustering. The process of geo-locating clusters is sketched in section 3. Subsequently, section 4 describes the structure of event extraction grammars, their creation, and multilinguality aspects. Further, section 5 elaborates on information fusion. The event visualization and accessing fully-fledged event descriptions generated by the system is presented in section 6. Some evaluation figures are given in section 7. Finally, we end with a summary in section 8.

1. Real-time Event Extraction Process

This section briefly describes the real-time event extraction processing chain, which is depicted in Figure 1. First, before the proper event extraction process can proceed, news articles are gathered by dedicated software for electronic media monitoring, namely the EMM system [1] that receives 50000 news articles from 1500 news sources in 41 lan-

[2] MUC - http://www.itl.nist.gov/iaui/894.02/related/projects/muc

[3] ACE - http://projects.ldc.upenn.edu/ace

guages each day. Secondly, all articles are classified according to around 700 categories and then scanned in order to identify known entities (e.g., geographical references, names of known people and organizations, etc.). This information is then created as metadata for each article. Next, the articles are grouped into news clusters according to content similarity. Subsequently, each cluster is geo-located. Further, clusters describing security-related events are selected via the application of key-word based heuristics.

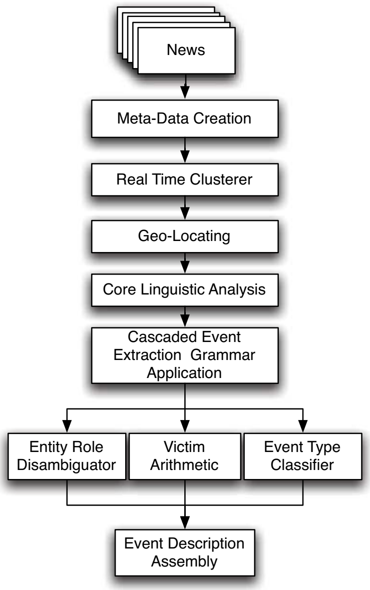


Figure 1. Real-time event extraction processing chain.

Next, each of these clusters is processed by NEXUS (News cluster Event eXtraction Using language Structures), the core event extraction engine. For each cluster it tries to detect and extract only the main event by analyzing all articles in the cluster. For each detected violent and natural disaster event NEXUS produces a frame, whose main slots are: date and location, number of killed and injured, kidnapped people, actors, and type of event. In an initial step, each article in the cluster is linguistically preprocessed in order to produce a more abstract representation of its text. This encompasses the following steps: fine-grained tokenization, sentence splitting, domain-specific dictionary look-up (e.g., recognizing numbers, quantifiers, person titles), labeling of key terms indicating unnamed person groups (e.g. *civilians*, *policemen*, *Shiite*), and morphological analysis. The aforementioned tasks are accomplished by CORLEONE (Core Linguistic Entity Online Extraction), our in-house core linguistic engine [5].

Once the linguistic preprocessing is complete, a cascade of extraction grammars is applied on each article within a cluster. The patterns for the extraction grammars are created through a blend of machine learning and knowledge-based techniques. Contrary to other approaches, the learning phase is done via exploiting clustered news, which intuitively guarantees better precision of the learned patterns. The extraction patterns are matched against the first sentence and the title of each article from the cluster. By

processing only the top sentence and the title, the system is more likely to capture facts about the most important event in the cluster.

Finally, since the information about events is scattered over different articles, the last step consists of cross-article cluster-level information fusion in order to produce fully-fledged event descriptions, i.e., we aggregate and validate information extracted locally from each single article in the same cluster. This process encompasses mainly three tasks, entity role disambiguation (as a result of extraction pattern application the same entity might be assigned different roles), victim counting and event type classification.

The core event-extraction process is synchronized with the real-time news article clustering system in EMM in order to keep up-to-date with most recent events. A more thorough description of the real-time news clustering, geo-locating clusters, event extraction grammars, information fusion, issues concerning multilinguality and accessing the results produced by the live event extraction engine follows in the subsequent sections.

2. Real-time Clustering of News Articles

The real time news clustering system performs periodical hierarchical clustering on a set of news articles harvested in a 4-hour time window. A cache is maintained for twice that period in case the number of articles drops below a certain minimum, in which case the system attempts to process at least that set minimum. This process is performed every 10 minutes.

Initially each news article is considered as a cluster. The clustering process is agglomerative and uses average group linkage to determine the distances between the clusters. For calculating these distances a simple cosine measure is used. The clustering process continues until the maximum cosine distance falls below a certain set threshold (a function of the theoretical density of the feature vectors, where a higher density leads to a higher threshold value).

The article feature vectors are simple word count vectors. The words used for the construction of the word-document space are selected based on various criteria depending on the amount of information available in the time window under consideration. No lemmatizing is performed and a simple bag-of-words approach is used instead.

The system maintains a constantly updated frequency table for all words found in all articles processed for a particular language over time. The system employs a technique similar to an infinite input response filter to calculate and maintain these frequencies, which can shift over time. After a fixed run-length the number of samples used to calculate the average is set. This prevents loss of accuracy, numeric overflow and gives a more realistic temporal view of the word frequencies. Additionally, a full information entropy calculation is performed over the article set in the time window every clustering run [6].

The selection criteria for the words to be used for the construction of the word-document space are as follows:

- only use words of more than 2 characters
- reject words that are in the 100 most frequently used words in the language
- reject words that have an information entropy higher than a certain language-dependent threshold (typically 0.75) in the document set under consideration
- consider only the first 200 words of each news article (length normalization)

- only select words that appear at least twice in at least 2 documents, unless this results in a number of unique words less than a certain threshold (typically 1000) in which case only use words that appear at least once in at least 2 documents

A cluster is only valid if it contains articles from at least 2 different sources and these articles are not duplicates.

After the clustering procedure the new clusters are compared with those calculated at $t - \Delta_t$. If there is any overlap between clusters, the articles appearing in cluster at $t - \Delta_t$ which do not appear in the current cluster, are linked to the current cluster. There are 2 possible reasons for the missing articles: either the articles have simply dropped out of the time window, or the cluster has broken up due to the changes in word-document space. This merging process can lead to internal duplication of articles, i.e. the articles are linked back into the original cluster, but also exist in a newly formed cluster. In order to avoid this duplication, the current cluster set is scanned for internal overlap and any overlapping clusters are merged.

In this way news stories consisting of large numbers of articles, far exceeding the number present in any time-window, can be tracked incrementally over time, as long as there is some reporting in the time-window.

3. Geo-Locating

Homographs pose a well-known problem in the process of geo-tagging news articles [7]. In particular, words referring to place names may: (a) occur as person names (e.g., *Conrad Hilton* are the names of towns in the USA, New Zealand and Canada); (b) occur as common words (e.g., *And* is a village in Iran); or (c) have variants, i.e., well known cities have language variants (e.g., *Seul* meaning alone/only in French is also the Capital of South Korea in Portuguese and Italian), (d) refer to different locations (e.g., there are 33 places named *Clinton*). Additional complications are caused by language inflection and capital cities referring to the reporting location.

Our latest approach is to conduct homograph disambiguation at two distinct levels, firstly at the level of individual article and secondly, when the articles have been clustered. The article geo-tagging algorithm occurs as follows: First, the problem in (a) is solved via removing from the text names of known people and organisations. Next, a multi-lingual gazetteer of place, province, region and country names is used to geo-match a list of candidate locations in the news articles.

In order to disambiguate homographs that are common words and place names (see (b) and (d)), the traditional approach is to use language dependent stop word lists. We use a different approach based on two characteristics maintained in our gazetteer. The first characteristic classifies locations based on their perceived size, such that capital cities and major cities have a higher class than small villages. The second characteristic maintains the hierarchical relation of place in its administrative located hierarchy (i.e., town, in province, in region, in country). The disambiguation algorithm lets high class locations pass through as well as locations that have a containment relation with other candidate locations (e.g., *Paris*, *Texas*, *USA*). To resolve (c) we remove any geo-tags that are incompatible with the language of the article.

In order to handle name inflection, names are maintained with their variants encoded as a regular expression. Only matches, for which the language of the name and the article

are compatible, are maintained in the candidate location list. Next, a Newswire location filtering is applied, where the word position of the candidate location is used to promote locations occurring after an initial location since the Newswire location generally appears earlier in the article.

At the level of the cluster all named entities and geo-tags in the articles of the cluster are gathered together. This is necessary to remove named entities that have not been matched in articles (a common reporting trait is to mention only well known names, e.g., *Clinton*, *Hilary Rodham Clinton*, *Hilary Rodham* and *Hilary Clinton*). Next, we gather the highest score for each place by maximum value (we don't sum same place values since often news wires are repeated many times and this would affect the real location that is cited in the article). Next, we cap the scores, here we look for tags that are either countries or regions and check whether any of the places cited are physically inside. When this is the case we increase the place score, i.e., some articles of the same cluster will talk more generally of an event in China while others will be more specific and will describe a location like Sechuan Province. Finally, the highest scoring place tag is chosen as the geo-location of the cluster.

Compared to the previous version of the geo-location algorithm [8] we have carried out a more thorough evaluation for English, French and Italian. The data set used were the daily clusters over three selected days. The results are presented in Table 1.

Table 1. Precision and Recall of the Geo-Location algorithm

Language	Cluster Size	Correct Tags	Missed Tags	False Positive	Precision	Recall
English	294	270	17	7	97%	94%
French	77	76	1	4	95%	98%
Italian	83	74	9	1	99%	89%

4. Cascaded Event Extraction Grammars

In order to detect event information from news articles we deploy a cascade of finite-state extraction grammars. Following this approach was mainly motivated by the fact that: (a) finite-state grammars can be efficiently processed , and (b) using cascades of smaller grammars facilitates the maintenance of underlying linguistic resources and extending the system to new domains.

This section first introduces our in-house IE-oriented pattern matching engine. Subsequently, the structure of the current event extraction grammar for English is presented. Finally, the automatic extraction pattern acquisition and issues concerning adapting event extraction process to new languages are briefly addressed.

4.1. Pattern Matching Engine

In order to guarantee that massive amounts of textual data can be digested in real time, we have developed ExPRESS (Extraction Pattern Engine and Specification Suite), a highly efficient extraction pattern engine [11], which is capable of matching thousands of patterns against MB-sized texts within seconds. The specification language for creating extraction patterns in ExPRESS is a blend of two previously introduced IE-oriented gram-

mar formalisms, namely JAPE (Java Annotation Pattern Engine) used in the widely-known GATE platform [9] and XTDL, a significantly more declarative and linguistically elegant formalism used in a lesser known SPROUT platform [10].

An EXPRESS grammar consists of pattern-action rules. The left-hand side (LHS) of a rule (the recognition part) is a regular expression over flat feature structures (FFS), i.e., non-recursive typed feature structures (TFS) without structure sharing, where features are string-valued and unlike in XTDL types are not ordered in a hierarchy. The right-hand side (RHS) of a rule (action part) constitutes a list of FFS, which will be returned in case LHS pattern is matched.

On the LHS of a rule variables can be tailored to the string-valued attributes in order to facilitate information transport into the RHS, etc. Further, like in XTDL, functional operators (FO) are allowed on the RHSs for forming slot values and for establishing contact with the 'outer world'. The predefined set of FOs can be extended through implementing an appropriate programming interface. FOs can also be deployed as boolean-valued predicates. The two aforementioned features make EXPRESS more amenable than JAPE since writing 'native code' on the RHS of rules (common practice in JAPE) has been eliminated. Finally, we adapted the JAPEs feature of associating patterns with multiple actions, i.e., producing multiple annotations (possibly nested) for a given text fragment. It is important to note that grammars can be cascaded. The following pattern for matching events, where one person is killed by another, illustrates the syntax.

```
killing-event :>
  ((person & [FULL-NAME: #name1]):killed
    key-phrase & [METHOD: #method, FORM: "passive"]
   (person & [FULL-NAME: #name2]):killer):event
  ->
  killed: victim & [NAME: #name1],
  killer: actor & [NAME: #name2],
  event: violence & [TYPE: "killing",
                     METHOD: #method,
                     ACTOR: #name2,
                     VICTIM: #name1,
                     ACTOR_IN_EVENTS: numEvents(#name2)]
```

The pattern matches a sequence consisting of: a structure of type `person` representing a person or group of persons who is (are) the victim, followed by a key phrase in passive form, which triggers a *killing event*, and another structure of type `person` representing the actor. The symbol `&` links a name of the FFS's type with a list of constraints (in form of attribute-value pairs) which have to be fulfilled. The variables `#name1` and `#name2` establish bindings to the names of both humans involved in the event. Analogously, the variable `#method` establishes binding to the method of killing delivered by the `key-phrase` structure. Further, there are three labels on the LHS (`killed`, `killer`, and `event`) which specify the start/end position of the annotation actions specified on the RHS. The first two actions (triggered by the labels `killed` and `killer`) on the RHS produce FFS of type `victim` and `actor` resp., where the value of the `NAME` slot is created via accessing the variables `#name1` and `#name2`. Finally, the third action produces an FFS of type `violence`. The value of the `ACTOR_IN_EVENTS` attribute is computed via a call to a FO `numEvents()` which contacts external knowledge base to retrieve the number of events the current actor was involved in the past.

A comprehensive overview of the techniques for compiling and processing grammars and the entire ExPRESS engine as well as the results of some experiments on ExPRESS run-time behaviour comparison with XTDL and JAPE is given in [11].

4.2. Event Extraction Grammars

There are two different approaches to building event extraction grammars: (i) the complexity of the language is represented at the level of lexical descriptions, where each word in the lexicon is provided with a rich set of semantic and syntactic features ([12] and [13]); (ii) the complexity of the language is represented through different, mostly linear, patterns in the grammar, which rely on superficial or less sophisticated linguistic features [14].

Providing rich lexical descriptions like verb sub-categorization frames in [12] or ontologies in [13] requires a linguistic expertize, on the other hand more shallow lexical descriptions will result in more patterns to encode the necessary linguistic knowledge. However, superficial patterns are closer to the text data. We believe that their creation is more intuitive and easier for non-experts than building ontologies or sub-categorization frames. Writing such patterns is easier for languages like English, where the word ordering obeys strict rules and the morphological variations are not an important issue. Therefore, we followed this approach when creating our English event extraction grammar.

This grammar in its current version consists of two subgrammars. The first-level subgrammar contains patterns for recognition of named-entities, e.g., person names (*Osama bin Laden*), unnamed person groups (*at least five civilians*), and named person groups (e.g., *More than thousands of Iraqis*). As an example consider the following rule for detecting mentions of person groups.

```
person-group :> ((gazetteer & [GTYPE: "numeral",
                               SURFACE: #quant,
                               AMOUNT: #num])
                 (gazetteer & [GTYPE: "person-modifier"])?
                 (gazetteer & [GTYPE: "person-group-proper-noun",
                               SURFACE: #name1])
                 (gazetteer & [GTYPE: "person-group-proper-noun",
                               SURFACE: #name2])):name
  -> name: person-group & [QUANTIFIER: #quant,
                           AMOUNT: #num,
                           TYPE: "UNNAMED",
                           NAME: #name,
                           RULE: "person-group"],
                        & #name = Concatenate(#name1,#name2).
```

This rule matches noun phrases (NP), which refer to people by mentioning their nationalities, religion or political group to which they belong, e.g. *three young Chinese*, *one Iraqi Muslim*, *three young Maoists*. The words and phrases which fall in the category `person-group-proper-noun`, `numeral` and `person-modifier` (e.g. *young*) are listed in a domain-specific lexicon (gazetteer). In this way the grammar rules are being kept language independent, i.e., rules can be applied to languages other than English, provided that language-specific dictionaries back up the grammar. Through abstracting from surface forms in the rules themselves the size of the grammars can be kept relatively low and any modifications boils down to extend-

ing the lexica, which makes the development for non-experts straightforward. Further, the first-level grammar does not rely on morphological information and uses circa 40 fine-grained token types (e.g., `word-with-hyphen`, `opening-bracket`, `word-with-apostrophe`, `all-capital-letters`), which are to a large extent language independent. Consequently, the majority of the first-level grammar rules for English can be used for processing texts in other languages. Clearly, some of these rules might not be applicable for some languages due to the differences in syntactic structure, but they are intuitively easily modifiable.

The second-level subgrammar consists of patterns for extracting partial information on events: actors, victims, type of event, etc. Since the event extraction system is intended to process news articles which refer to security and crisis events, the second-level grammar models only domain-specific language constructions. Moreover, the event extraction system processes news clusters which contain articles about the same topic from many sources, which refer to the same event description with different linguistic expressions. This redundancy mitigates the effect of phenomena like anaphora, ellipsis and complex syntactic constructions. As mentioned earlier, the system is processing only the first sentence and the title of each article, where the main facts are summarized in a straightforward manner, usually without using coreference, sub-ordinated sentences and structurally complex phrases. Therefore, the second-level grammar models solely simple syntactic constructions via utilization of 1/2-slot extraction patterns like the ones given below. The role assignments are given in brackets.

```
[1] PERSON-GROUP <DEAD> "were killed"
[2] {PERSON | PERSON-GROUP} <DEAD> "may have perished"
[3] PERSON-GROUP <DEAD> "dead"
[4] "police nabbed" PERSON <ARRESTED>
[5] PERSON-GROUP <DISPLACED> "fled their homes"
[6] PERSON <DEAD> "was shot dead by" PERSON <PERPETRATOR>
```

These patterns are similar in spirit to the ones used in AutoSlog [15]. They are acquired semi-automatically (see 4.3). The fact that in English the word ordering is more strict and the morphology is simpler than in other languages contributes also to the coverage and accuracy of the patterns, which encode non-sophisticated event-description phrases.

To sum up, event extraction system discards the text, which goes beyond the first sentence for the following reasons: (a) handling more complex language phenomena is hard and might require knowledge-intensive processing, (b) the most crucial information we seek for is included in the title or first sentence, (c) if some crucial information has not been captured from one article in the cluster we might extract it from other article in the same cluster.

In order to keep the grammar concise and as much as possible language independent, we represent in the grammar all surface-level linear patterns via pattern types, which indicate the position of the pattern with respect to the slot to be filled (left or right). To be more precise, all patterns are stored in a domain-specific lexicon (applied prior to grammar application), where surface patterns are associated with their type, the event-specific semantic role assigned to the entity which fills the slot (e.g., `WOUNDED`, `KIDNAPPED`, `DISPLACED`, etc.) and the number of the phrase which may fill the slot (singular, plural, or both). For instance, the surface pattern `"shot to death"` for recognizing dead victims as a result of shooting, is encoded as follows.

```
shot to death [TYPE: right-context-sg-and-pl,
                SURFACE: "shot to death",
                SLOTTYPE: DEAD]
```

The value of the `TYPE` attribute indicates that the pattern is on the right-hand-side of its slot (`right-context`), which can be filled by a phrase which refer to one or many people (`sg-and-pl`). The event-specific semantic role, assigned to each NP filling the slot is `DEAD`. Via such an encoding of event-triggering linear patterns, the extraction patterns [1] and [3] from the list given above are merged into one (in a simplified form):

```
dead-person :> person-group
               gazetteer & [TYPE: right-context-sg-and-pl SLOT: dead]
      -> ...
```

Interestingly, we have found circa 3000 event-triggering surface patterns for English. Clearly, the strict word ordering and relatively simple morphology allowed for easy generation of pattern variants.

4.3. Pattern Acquisition

For creating second-level grammar patterns described in previous section a weakly supervised machine learning (ML) algorithm has been deployed. It is similar in spirit to the bootstrapping algorithms described in [16,17]. In particular, the pattern acquisition process involves multiple consecutive iterations of ML followed by manual validation. Learning patterns for each event-specific semantic role requires a separate cycle of learning iterations. The method uses clusters of news articles produced by EMM. Each cluster includes articles from different sources about the same news story. Therefore, we assume that each entity appears in the same semantic role in the context of one cluster ('one sense per discourse'). The core steps of the pattern acquisition algorithm are as follows:

1. Annotate a small corpus with event-specific information
2. Learn automatically single-slot extraction patterns from annotated corpus
3. Manually validate/modify these patterns
4. If the size of the pattern set exceeds certain threshold, then terminate
5. Match the patterns against the full corpus or part of it
6. Assign semantic roles to entities which fill the slots
7. Annotate automatically all the occurrences of these entities in all articles in the same cluster with the corresponding role this entity has been assigned in the previous step and goto 2

An automatic procedure for syntactic expansion complements the learning, i.e., based on a manually provided list of words which have identical (or nearly identical) syntactic model of use (e.g. *killed*, *assassinated*, *murdered*, etc.) new patterns are generated from the old ones by substituting for each other the words in the list. Subsequently, some of the automatically acquired single-slot patterns were used to manually create 2-slot patterns like `X shot Y`. The pattern acquisition process is described thoroughly in [18].

4.4. Multilinguality

The English event extraction grammar described in the previous section relies mainly on a multilingual named-entity recognition grammar, language-specific dictionary of NE-

relevant trigger words, and a language-specific dictionary of surface-level event extraction patterns. In this way adapting the current event extraction grammar to new languages does not require linguistic expertize.

We have extended the system to the processing of Italian and French in a brute-force manner via providing the aforementioned language-specific resources. Although the system's precision was high as expected, an empirical coverage analysis for Italian showed that that there are some drawbacks when surface-level patterns are applied. This is mainly due to the free-word order and relative morphological richness of Romance languages. Further, Italian appears to be more verbose with the respect to expressing information about events, i.e., it is structurally more complex. Consequently, we designed a slightly more linguistically sophisticated rules for Italian, which cover a rich variety of morphological and syntactic constructions. The details of the process of adapting the system to the processing of Italian are presented more thoroughly in [19].

Although we intend to develop linguistic resources for tackling the event extraction task for Arabic (EMM has some 86 different news sources in this language) and apply simmilar techniques as for English, our first approach is slightly different. First, statistical machine translation systems are used for translating the articles in the Arabic clusters into English. Next, these translations are passed into the English event extraction system. In particular, we integrated two Arabic-to-English translation systems into the live processing chain, namely *Google Translate*[4] and *LanguageWeaver*[5]. Currently we are evaluating the results of this approach.

5. Information Fusion

Once the event extraction grammars has been applied locally at document level the single pieces of information are merged into fully-fledged event descriptions. In particular, three tasks are performed at this level: (a) semantic role disambiguation, (b) victim counting, and (c) event type classification. They are described briefly.

Semantic Role Disambiguation: If one and the same entity has two roles assigned in the same cluster, a preference is given to the role assigned by the most reliable group of patterns. The double-slot patterns are considered the most reliable. Regarding the one-slot constructions, patterns for detection of *killed*, *wounded*, and *kidnapped* are considered as more reliable than the ones for extraction of the *actor (perpetrator)* slot (the latter one being more generic). The pattern reliability ranking is based on empirical observations.

Victim Counting: Another ambiguity arises from the contradictory information which news sources give about the number of victims (e.g., killed) An ad-hoc algorithm for computing the most probable estimation for these numbers is applied. It finds the largest group of numbers which are close to each other and subsequently finds the number closest to their average. All articles, which report on number of victims which significantly differs from the estimated cluster-level victim number are discarded. In order to perform victim arithmetics at the document level a small taxonomy of person classes is used.

[4]http://translate.google.com/translate_t?hl=en
[5]http://www.languageweaver.com

Event Type Classification: The event type classification algorithm uses a blend of keyword matching and domain specific rules. First, all potential event types are assigned ranks based on the occurrence of type-specific keywords in the articles in a given cluster (a rank for a given type is set to zero if no keywords related to this type were found in the articles in the given cluster). Next, a non-zero valued rank of a more specific event type[6] is boosted in case the rank of the type which subsumes it is non zero. Finally, the type with the highest rank is selected, unless some domain-specific event type classification rule (they have higher precedence) can be applied. As an example, consider the following domain specific rule: if the event description includes named entities, which are assigned the semantic role *kidnapped*, as well as entities which are assigned the semantic role *released*, then the type of the event is *Hostage Release*, rather than *Kidnapping*.

6. Live Event Tracking

Multilingual crisis-related event tracking poses a number of practical issues, mainly related to the correct geo-spatial visualization of the event together with it's principal characteristics. Another concern is to minimize constraints on end users to rely on expensive and proprietary desktop applications. We fulfill these issues by publishing the event data using current internet standard formats, namely, KML and GeoRSS. In particular, the results of the event extraction are accessible in two ways: (a) via *Google Earth* application which is passed event descriptions in KML format[7]; and (b) via a publicly accessible web client[8] that exploits the *Google Maps* technologies and connects to our KML server.

The event description interface provides elements in two languages. The title, description and the precise location of the event is presented in the native language of the event, i.e., the language of the cluster being processed. In the English language we provide the event consequences and the relative geographical path to the place. The diagram in Figure 2 shows the user interface for French, Arabic and Italian respectively.

To provide immediate clues on the cause and effect of the event we use three visual indicators: the icon image of the event type that is geo-locted on the map; the size of the icon depends on the size of the cluster; and the magnitude of the consequence of the event is indicated by a colored circle around the event. A key to the symbols used in our system is given in Figure 3.

7. Event Extraction Evaluation

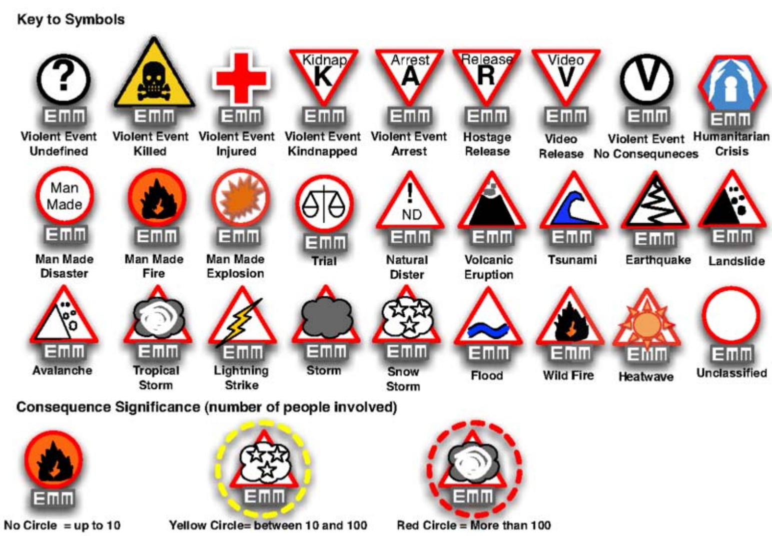
An evaluation of the event extraction performance has been carried out on 368 English-language news clusters based on news articles downloaded on 24 January 2008. 29 violent events are described in these clusters and 27 out them were detected by our system (**93% coverage**). In some cases several clusters referred to the same event. We consider that an event is detected by the system, if at least one cluster referring to it was captured.

[6]Types are ordered in a event type hierarchy

[7]For English: start Google Earth with KML: http://press.jrc.it/geo?type=event&format=kml&language=en. For other languages change the value of the language attribute accordingly.

[8]http://press.jrc.it/geo?type=event&format=html&language=en

Figure 2. Event visualization in Google Earth: Diagram showing events detected in French, Arabic and Italian. Note that the event meta data is expressed in English whilst the original, title and description are maintained.



Our system detected 55 news clusters which refer to 37 violent and non-violent events. 27 out of these 37 detected events are violent events (**73% precision**). In the context of crisis monitoring, discovery of disasters causing victims may also be considered relevant. Our system detected 3 man made disasters; if we consider them relevant together with the violent events, then the precision of the event extraction becomes **81%**. With the respect to victim counting, for the number of dead and injured an accuracy of 80% and 93% could be achieved respectively. The event type detection performs significantly worse, i.e., only in 57% of the cases the correct event type could be assigned.

One of the most frequent errors in event classification was in case of clusters referring to terrorist bombing, which were not classified as *Terrorist Attack*, but simply as a *Bombing*. Another problem is that classification based on simple keyword matching sometimes gives wrong results, e.g., an airplane crash was classified as *Air Attack*, since a military plane was involved. Further, there are situations, in which two different incidents are reported as a part of one event. For example, a terrorist attack might include bombing and shooting. In such cases only one label is assigned to the whole event, although two labels would be more appropriate.

We have repeated the evaluation of the event extraction on several days, which yielded on an average similar accuracy figures to the ones reported above.

8. Summary

This paper presented a real-time and multilingual news event extraction system.[9] Currently, it is capable of efficiently processing news in English, Italian, French and Arabic, which provide descriptions of the latest crisis-related events around the world with a 10-minute delay. The results of the evaluation on violent and natural disaster event extraction show satisfactory performance and application potential for real-time global crisis monitoring. In order to guarantee ease in maintenance and extensibility to new languages and domains a shallow text processing strategy has been chosen, which utilizes clustered data at different strata. The results of the live event extraction are provided via a publicly accessible web page and can be also accessed with the *Google Earth* application.

In order to improve the quality of the extracted event descriptions, several system extensions are envisaged. Firstly, we are working towards fine grained classification of natural and man made disasters. We also plan to improve the classification accuracy for the violent events. Secondly, we aim at extending the system to new languages. This is feasible, since our algorithms and grammars are mostly language independent. In particular, we currently focus on adapting the system to the processing of Russian, Spanish, Polish, Portuguese and Arabic news. Being able to extract information about same event in different languages and cross-linking them adds an additional navigation layer and an opportunity for fact validation and improving the accuracy of the system. Therefore, we plan to explore techniques for refining the results of event extraction, which are similar in spirit to those presented in [20]. Furthermore, we envisage to study the usability of event descriptions collected over time for the same purpose. Finally, a long-term goal is

[9]The work presented in this paper was supported by the Europe Media Monitoring (EMM) Project carried out by the Web Mining and Intelligence Action in the Joint Research Centre of the European Commission. We are indebted to all our EMM colleagues without whom the presented work could not have been be possible.

to automatically discover structure of events and relations between them, i.e., discovering sub-events or related events of the main one.

References

[1] C. Best, E. van der Goot, K. Blackler, T. Garcia, and D. Horby. Europe Media Monitor. Technical Report EUR 22173 EN, European Commission, 2005.

[2] R. Grishman, S. Huttunen, and R. Yangarber. Real-time Event Extraction for Infectious Disease Outbreaks. In *Proceedings of Human Language Technology Conference 2002*, San Diego, USA, 2002.

[3] G. King and W. Lowe. An Automated Information Extraction Tool For International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design. *International Organization*, 57:617–642, 2003.

[4] N. Ashish, D. Appelt, D. Freitag, and D. Zelenko. *Proceedings of the workshop on Event Extraction and Synthesis*, held in conjunction with the AAAI 2006 conference, Menlo Park, California, USA. 2006.

[5] J. Piskorski. CORLEONE – Core Linguistic Entity Online Extraction. Technical report EN 23393, Joint Research Center of the European Commission, Ispra, Italy, 2008.

[6] C. Shannon. A mathematical theory of communication. In *The Bell System Technical Journal*, Vol. 27, 1948.

[7] B. Pouliquen, M. Kimler, R. Steinberger, C. Ignat, T. Oellinger, K. Blackler, F. Fuart, W. Zaghouani, A. Widiger, A. Forslund, and C. Best. Geocoding multilingual texts: Recognition, Disambiguation and Visualisation. In *Proceedings of LREC 2006*, Genoa, Italy, pages 24–26, 2006.

[8] H. Tanev, J. Piskorski, M. Atkinson. Real-Time News Event Extraction for Global Crisis Monitoring. In *Proceedings of the* 13^{th} *International Conference on Applications of Natural Language to Information Systems (NLDB 2008).* Lecture Notes in Computer Science Vol. 5039, Springer-Verlag Berlin Heidelberg, pages 207–218, 2008.

[9] H. Cunningham, D. Maynard, and V. Tablan. Jape: a Java Annotation Patterns Engine (second edition). Technical Report, CS–00–10, University of Sheffield, Department of Computer Science, 2000.

[10] W. Drożdżyński, H.-U. Krieger, J. Piskorski, U. Schäfer, and F. Xu. Shallow Processing with Unification and Typed Feature Structures — Foundations and Applications. *Künstliche Intelligenz*, Vol. 1, 2004.

[11] J. Piskorski. ExPRESS – Extraction Pattern Recognition Engine and Specification Suite. In *Proceedings of the International Workshop Finite-State Methods and Natural language Processing 2007 (FSMNLP'2007)*, Potsdam, Germany, 2007.

[12] C. Aone, M. Santacruz. REES: A Large-Scale Relation and Event Extraction System. In *Proceedings of ANLP 2000,* 6^{th} *Applied Natural Language Processing Conference*, Seattle, Washington, USA, 2000

[13] B. Popov, A. Kiryakov, D. Ognyanoff, D. Manov, A. Kirilov, M. Goranov. Towards Semantic Web Information Extraction. In *Proceedings of International Semantic Web Conference*, Sundial Resort, Florida, USA, 2003.

[14] R. Yangarber, R. Grishman. Machine Learning of Extraction Patterns from Un-annotated Corpora. In *Proceedings of the* 14^{th} *European Conference on Artificial Intelligence: Workshop on Machine Learning for Information Extraction*, Berlin, Germany, 2000.

[15] E. Riloff. Automatically Constructing a Dictionary for Information Extraction Tasks. In *Proceedings of the* 11^{th} *National Conference on Artificial Intelligence (AAAI-93)*, , MIT Press, 1993.

[16] R. Jones, A. McCallum, K. Nigam, and E. Riloff. Bootstrapping for Text Learning Tasks. In *Proceedings of IJCAI-99 Workshop on Text Mining*, Stockholm, Sweden, 1999.

[17] R. Yangarber. Counter-Training in Discovery of Semantic Patterns. In *Proceedings of the* 41^{st} *Annual Meeting of the ACL*, 2003.

[18] H. Tanev and P. Oezden-Wennerberg. Learning to Populate an Ontology of Violent Events (in print). In F. Fogelman-Soulie, D. Perrotta, J. Piskorski, and R. Steinberger (editors), *NATO Security through Science Series: Mining Massive Datasets for Security*. IOS Press, 2008.

[19] V. Zavarella, J. Piskorski, H. Tanev. Event Extraction for Italian using a Cascade of Finite-State Grammars. In *Proceedings of the 7th International Workshop on Finite-State Machines and Natural Language Processsing*, Ispra, Italy, 2008.

[20] H. Ji, R. Grishman. Refining Event Extraction through Unsupervised Cross-document Inference. In *Proceedings of* 46^{th} *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Columbus, Ohio, USA, 2008.

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-291

Adaptive System for the Integration of Recommendation Methods with Social Filtering Enhancement

Przemysław KAZIENKO and Paweł KOŁODZIEJSKI
Technical University of Wrocław, Institute for Applied Informatics,
Wybrzeże Wyspiańskiego 27, 53-370 Wrocław, Poland
e-mail: przemyslaw.kazienko@pwr.wroc.pl

Abstract. We try to find the new application for recommender systems: online social networks. The innovative system for personalized and adaptive integration of recommendation methods is presented. The unique idea of the social filtering for network members is introduced and incorporated into recommender system.

Keywords. recommender system, social filtering, integrated systems, adaptation

Introduction

Recommender systems are widely used in web-based systems, especially in e-commerce [2], [3], [9], [13], [14]. They can fulfil user needs in many ways [8]. However, some application domain as social networks are unique and their special characteristics must be taken into consideration, resulting in a new approach to recommendation [1], [6].

Social networks have two goals: to group as many people together as possible and to create many strong relationships between them. Attachment of new users helps network grow and is the best way to introduce new ideas to the users and to keep community alive. Creating many strong relationships with a user builds up locality and prevents members from leaving.

The first challenge to consider is to attract new users to join the network and recommendations are the good way to reach it. If a new user or guest will become interested in recommended members, the network would win the new serious member. Another problem is how to recommend somebody to a new user having not much information about him/her. Another difficulty poses the recommendations to existing users [15].

There are two approaches, supporting either strong or weak ties in the network. The strength of a tie between two members is proportional to the emotional effort they invested in their relationship. A strong tie exists between members of a family and a week one between people that saw each other once in their lives. Supporting strong ties promotes small, close networks with a high loyalty.

Weak ties create big, widespread networks that grow fast. Their members are constantly exposed to new ideas making them very interesting. Because of their low loyalty level, members with weak links can easily leave our network in favour of

another one. This suggests that the best approach would be a hybrid of both solutions, to create a network that first bounds users with strong ties to prevent them from leaving the system, and next, it promotes weak ties to keep user interested.

One of the first methods for recommendation of humans based on social network analysis (SNA) in the telecommunication social networks was presented in [7].

The recommendations that enhance willingness to meet other people who visit the same places and in this way belong to the same informal group was studied in [5]. It was based on localization and identification of informal groups of people.

Social networks were also used to support recommendation of users based on anthologies processing [11]. The other approaches utilizing anthologies have been developed in [12], [16].

1. Recommender System with Social Filtering Enhancement

1.1. General Concept, Component Recommendation Methods

The main concept of proposed method is to overcome shortcomings of a single recommendation method and to deliver full personalization [14], which could offer every user a separate list of potential friends. It simultaneously depends on navigation path, history of user's behaviour (e.g. ratings), and user's likes and dislikes as well as on effectiveness of previous recommendations for the given user.

To achieve full personalization, the system combines association rules for ephemeral content-based personalization as well as collaborative and demographic filtering for persistent one.Consequently, a complete hybrid recommender system is obtained that integrates many independent recommendation methods in personalized and adaptive way. It exploits weights that are dynamically recalculated according to the effectiveness of recommendation: the more effective the particular method, the bigger the weight it will gain. Every user has his or her own set of weights corresponding to the method's usefulness for this individual. The system also uses its knowledge achieved from previous users to better suit its new users. At the first launch, all weights are set to system initial base values. After some users join the system, its base weights are recalculated. Every new user starts from the current system base weights as initial values. Once a user gets his own weights set, only his personal behaviour has an influence on it.

The work of the recommender system starts with the user's interaction (Fig. 1). The context of interaction (the web page and user ID) determines which conditions have been fulfilled and which methods are allowed to present their recommendation lists.

The system is capable of integrating any number of independent methods, although only five have been used in the project. To attract new guests into the system, simple statistic method (method no. 1) is used. Others can rate every user, and those with the highest score are presented.

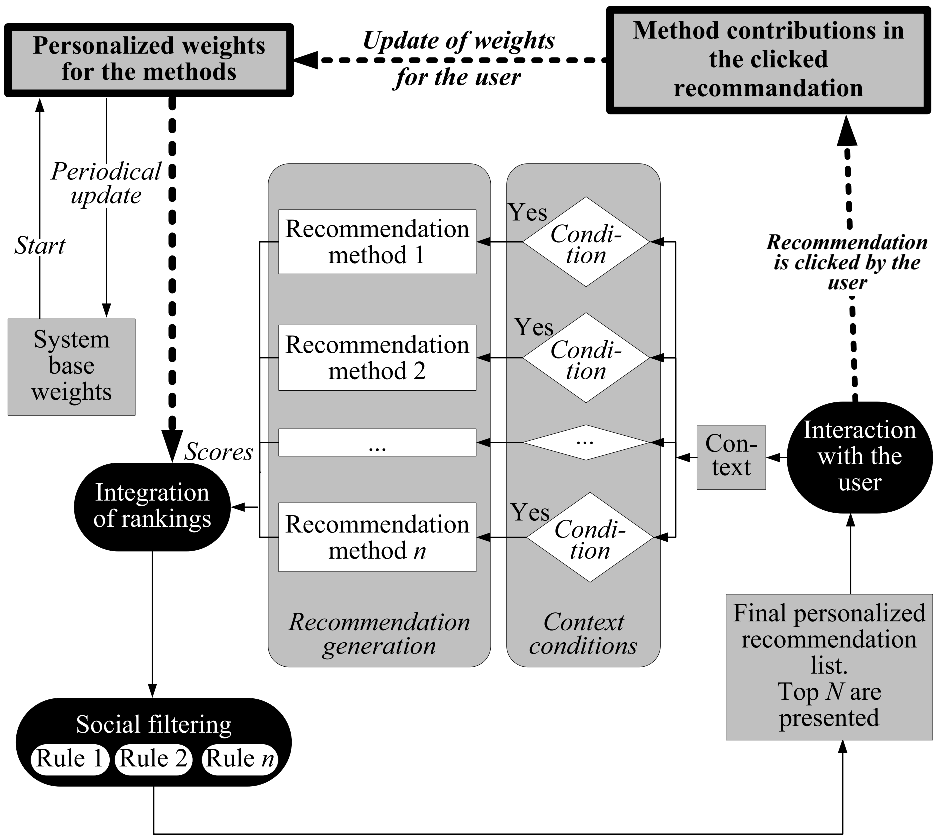


Figure 1. The concept of personalized integration of various recommendation methods. Dotted arrows correspond to the basic adaptation loop

As soon as user joins the system, the demographic method (method no. 2) can be used. It compares members based on the questionnaires that they filled up while creating their accounts.

If the purpose of the network is to spread it around the world, the system will score higher the users living far away from the given one rather than neighbours.

After some interaction with a new member, the system can use association rules (method no. 3) to find other users who navigated in the similar way.

Collaborative filtering (method no. 4) looks for members with a similar taste. It can be used for users, who rated any content: someone's images, blogs, or even actual users.

The last method, content based filtering (method no. 5) indexes all text input from users: blogs, "about me", even search queries, and finds users that employ the same keywords.

1.2. Method Integration

Every method relays its own list of recommended members with assigned, only positive values of scores. The system integrates, normalizes, and orders the received lists using both the obtained scores and weight set that belongs to the given user:

$$f_{jkl} = \sum_{i=1}^{M} \frac{w_{ik} * s_{ijkl}}{max_{ikl}}, \quad s_{ijkl} \geq 0 \qquad \textbf{(1)}$$

where:

f_{jkl} – final score of recommended member j for user k in context l;
w_{ik} – current weight of method i for user k;
s_{ijkl} – score of user j assigned by method i for user k with respect to context l;
M – number of methods,
max_{ikl} – maximum value of score s_{ijkl} among scores returned by method i – the top one in the i-th ranking.

Some discussion about the integration concept can be found in [9, 10].

1.3. Social Filtering

After integrating results from all methods into one list, system performs social filtering specialized for social networks. It is based on behaviour patterns observed by the social science and it should be adjusted to fit the needs of particular system.

Social filtering consists of a set of independent rules applied sequentially, rule by rule. The order of the rules is fixed and established by the system designers.

The goal of the first rule of social filtering is to enhance the exchange of ideas in the network by joining groups together. Users have tendency to create close clusters in which everyone knows one another. Those clusters are good to build members loyalty, but they do not introduce any new ideas. Hence, a good approach is to join those clusters together, and stimulate the information exchange. It is achieved in two steps. First, the system uses clustering algorithms to obtain groups of members based on relationships between users. A relationship is characterized by its existence, duration, and the intensity of mutual contacts like chats, emails, watching photos, blogs, etc.

Next, the filter rejects all users who belong to the same group with the given user. If two users from different groups will become friends, they are a first link between those two groups.

Another element (rule) of social filtering is related to the Dunbar's number of maximum connections: a person can keep track of up to 150 human relationships [4]. Thus, there is no point in recommending anybody to someone with 150 or more friends. Moreover, people with the small number of connections are preferred.

Other rules of filtering depend on the network policy, e.g. a physical distance between members. Some networks might aspire to become worldwide, so they recommend users who are most distanced from each other. In opposite, since some others might choose more local approach, they recommend users who are close ones to another hoping, that they will create a stronger relationship.

Yet another, basic, social filter is a relevancy filter. It rejects senseless recommendations like suggestions of members with no common language.

Note that it may happen that the output list of humans for recommendation after the filtering process can be very short or even empty. However, it appears that it is better not to suggest anybody rather than recommend irrelevant users. The latter case can annoy users being suggested.

1.4. Recommendation Presentation

The top N candidates from the final, filtered recommendation list are presented to the user. Obviously, in case of only few other users who match the current one, this list can be shorter or even empty.

Additionally, the system stores component scores for each item displayed to the user until the next user's contact. It enables to make use of latest user activities at estimation of recent recommendations.

1.5. Adaptation – Making Use of User Feedback

If a user chooses one of the recommendations, that links to another member j, the system checks what score s_{ijk} had each i-th method in recommending this member j. In consequence, the system adequately updates the appropriate weights of all methods in the set maintained for user k, as follows:

$$w'^{(1)}_{ik} = w^{(0)}_{i} + s_{ijkl} / max_{ikl} \text{, after the first click on recommendation performed by user } k$$

$$w'^{(n+1)}_{ik} = w^{(n)}_{ik} + s_{ijkl} / max_{ikl} \text{, after the } (n+1)\text{-th click} \quad \textbf{(2)}$$

where:

$w'^{(1)}_{ik}$, $w^{(n)}_{ik}$, $w'^{(n+1)}_{ik}$ – the weight of method i for user k after the first, n-th and n+1 user click on recommendation, respectively;

$w^{(0)}_{i}$ – the initial system base weight for method i.

An additional normalization mechanism is used to preserve the constant sum of weights:

$$w^{(n+1)}_{ik} = w'^{(n+1)}_{ik} * \frac{\sum_{i=1}^{M} w^{(n)}_{ik}}{\sum_{i=1}^{M} w'^{(n+1)}_{ik}} \quad \textbf{(3)}$$

where:

$w^{(n+1)}_{ik}$ is the normalized weight of method i for user k after the n+1 user click.

At the next click, $w^{(n+1)}_{ik}$ is used at calculation of $w'^{(n+2)}_{ik}$ with (2).

Note that, the initial system base weight can help to adjust the recommendation framework to the specific application domain. In an environment with the relatively big number of clicks on recommendation and general large activity of users, the greater system initial values can be more adequate and the single behaviour of a user would have rather smaller influence on their actual weights.

Similarly, if a typical user clicks on recommendation rather occasionally, the system should exploit this rare feedback more extensively by using smaller values of initial system base weights.

2. Conclusions and Future Work

The system successfully integrates five different recommendation methods. Those methods combined deliver the full time, personalized persistent recommendation. The additional social filter appears to be the key to success of recommender system, especially in the online social networks. It grasps the unique, immeasurable character of relationships between people. The careful adjustment of rules within that filter can stimulate members to tighten their relationships increasing loyalty or connect with new users for exchange of new ideas.

The future work should be concentrated on development of social filters in the close co-operation with sociologists.

Acknowledgments. The work was supported by The Polish Ministry of Science and Higher Education, grant no. N516 037 31/3708.

References

[1] L.A. Adamic, E. Adar, Friends and Neighbors on the Web. *Social Networks* 25 (2003), 211-230.
[2] G. Adomavicius, A. Tuzhilin, Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering* 17 (2005), 734-749.
[3] R. Burke, Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction* 12 (2002), 331-370.
[4] R.I.M. Dunbar, Co-evolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences* 16 (1993), 681-735.
[5] A. Gupta, S. Paul, Q. Jones, C. Borcea, Automatic identification of informal social groups and places for geo-social recommendations. *International Journal of Mobile Network Design and Innovation* 2 (2007), 159-171.
[6] J. Karamon, Y. Matsuo, M. Ishizuka, Generating Useful Network Features from Social Networks for Web 2.0 services. *WWW2008*.
[7] P. Kazienko, Expansion of Telecommunication Social Networks. The fourth International Conference on Cooperative Design, Visualization and Engineering, CDVE 2007, September 16-20, 2007, Shanghai, China, Lecture Notes in Computer Science LNCS 4674 (2007), Springer Verlag, 404-412.
[8] P. Kazienko, Web-based Recommender Systems and User Needs – the Comprehensive View. In: *New Trends in Multimedia & Network Information Systems*, IOS Press 2008 (to appear).
[9] P. Kazienko, P. Kołodziejski, Personalized Integration of Recommendation Methods for E-commerce. International Journal of Computational Intelligence 3 (2006), 12-26.
[10] P. Kazienko, P. Kołodziejski, WindOwls – Adaptive System for the Integration of Recommendation Methods in E-commerce. Third Atlantic Web Intelligence Conference AWIC 2005, Łódź, Poland, June 6-9, 2005, Lecture Notes in Computer Science LNAI 3528 (2005), Springer Verlag, 218-224.
[11] P. Kazienko, K. Musiał, T. Kajdanowicz, Ontology-based Recommendation in Multimedia Sharing Systems. 14th Americas Conference on Information Systems, AMCIS 2008, Toronto, Ontario, Canada, August 2008 (in press).
[12] S. Middleton, N. Shadbolt, D. De Roure, Ontological User Profiling in Recommender Systems. ACM Transactions on Information Systems, 22 (2004), 54–88.
[13] M. Montaner, B. López, J.L. de la Rosa, A Taxonomy of Recommender Agents on the Internet. Artificial Intelligence Review 19 (2003), 285-330.
[14] J.B. Schafer, J.A. Konstan, J. Riedl, E-Commerce Recommendation Applications. Data Mining and Knowledge Discovery 5 (2001), 115-153.
[15] E. Spertus, M. Sahami, O. Buyukkokten: Evaluating similarity measures: a large-scale study in the Orkut social network, *The eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD 2005, ACM Press, 2005, 678-684.
[16] C. Ziegler, W. Lindner, Semantic Web Recommender Systems, *9th International Conference on Extending Database Technology, Advances in Database Technology*, EDBT 2004, Lecture Notes in Computer Science LNCS 2992 (2004), Springer Verlag, 78–89.

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-904-2-297

Evaluation of Internet Connections

Andrzej SIEMIŃSKI
Wrocław University of Technology, Institute of Applied Informatics,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
e-mail: andrzej.sieminski@pwr.wroc.pl

Abstract. The user perceived latency of displaying Internet pages is one of main factors in his/her assessment of the network. There are many factors that have an influence on the latency. The paper studies one of them - the Internet connection provided by an ISP. Up to now, there is no comprehensive theoretical model of the relationship. Therefore, an experiment was necessary. During the experiment the latency of a variety of Web pages was measured. The results indicate that a popular assumption that tightly binds network throughput with browser latency is not well justified. The analysis of low level network connection properties shows a significant level of correlation between the duration of pages download and the trace route length between the browser and the source server. The trace routes to a given host recorded for a single ISP have a considerable, stable number of common hosts. Therefore, the analysis of trace routes can be used in the process of selection of an ISP.

Keywords. browser latency, network throughput, ISP, trace route

Introduction

No one likes to wait. This applies also to Internet surfing. The studies on human cognition [1], [2], [3] revealed that on the average a man easily accepts a latency lower than 10 seconds. The heavily cited study by Zona Research [4] estimated that each year Web sites lose several billions of dollars due to what they call "unacceptable download speed and resulting user bailout behavior".

There are many factors that influence the latency. The effects of caching on browser latency and of the so called page cacheability factor are discussed in depth in [5] and [6]. Loading at least a portion of objects from a local or proxy cache can speed the browsing considerably. Another factor is the technique of incremental loading of items. A good programmer of Web pages can fool the person into thinking a page is faster than what actually is by loading first the text and then including secondary menu graphics or high resolution graphics, and others. This gives the user something to start scanning and changes the perceptions of how fast a page loads. One can use also page compression and decompression to speed up browsing as described in [7].

All the above techniques are useful but a user has hardly any influence upon them. On the other hand he/she has a number of ISPs (Internet Service Providers) to choose from. The selection of an ISP depends on a variety of factors, most notably the reliability of the service, the cost, and the bandwidth. According to a general belief increasing the bandwidth should result in a proportional shortening the browser latency. The precise nature if the relationship is difficult to model. The paper tries to identify

the properties of Internet connections that most influence the browsing speed. The Internet connections provided by three ISPs in Poland have been tested.

The paper is organized as follows. The next section describes the adverse effects that some properties the Internet protocols have on the browsing speed. They are responsible for the fact, that increasing the ISP bandwidth does not necessary result in a corresponding increase in browsing speed. The deficiencies could be mitigated but not eliminated. Due to the complexity and heterogeneity of the Internet the analytical models are inadequate for both researchers and engineers [8]. Therefore, an experiment was necessary to gather data on the relationship between the properties of ISP offered Internet connection and the page download time. The third section summarizes its results. During the experiment several pages from different servers were periodically loaded and the download time was recorded. The measurements were done for three ISPs. At the same time several low level properties of the connections were also recorded. Sections 4 and 5 present the obtained results together with their statistical analysis. The section 6 studies the correlation between the browser latency and low level connection properties. The conclusions should help in the selection of an ISP. The future research work areas are discussed in the last 7th Section.

1. Mechanics of Internet Traffic

The TCP (Transmission Control Protocol) is used by the Web to deliver requests for objects from users to servers and Web objects in the opposite direction. The following TCP features have a profound effect on the performance of the Web:

- TCP ensures that lost packets of data are retransmitted;
- TCP is as a connection-oriented protocol;
- TCP regulates the rate the sending host transmits data with so the receiver can absorb it.

It is well known that the protocol was designed to maximize the reliability of data delivery not the speed of transmission. The protocol a notorious slow-starter, small objects not matter whether they have size of a few hundred or a few Kbytes are transmitted in a roughly the same time [9]. The reasons of the phenomena are twofold: firstly, the overhead of the transmission staring does not depend on the size of an object, and, secondly, the protocol adopts the size of transmitted packets of data to the current network load. Therefore, the network congestion is avoided at the cost of the transfer rate. The property is rather unfortunate as the Web traffic is characterized by the burst-outs of numerous requests for small objects.

The need to open and maintain a connection slows down the transmission. The message exchange to start a transaction (a three way handshake, see Fig. 1) is necessary because the TCP uses the unreliable IP protocol to transfer segments. The obvious inefficiency of the solution prompted the introduction of the persistent connections in the new HTTP/1.1 version of the protocol. The idea is to use the same TCP connection to send and receive multiple HTTP requests/responses thus reducing both network congestion (fewer TCP connections are made) and latency in subsequent requests (no handshaking). Modifying the infrastructure of a network spread all over the world is not an easy task – still many servers use the previous versions of the protocol. Initial report that analysis the extent to which different browsers are capable of exploiting the persistent connection is described in [10].

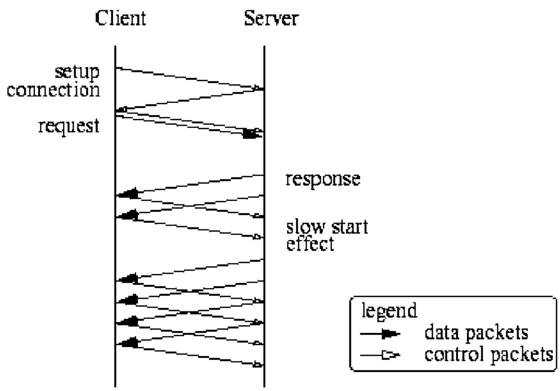


Figure 1. Establishing a HTTP connection between a client and a server [9]

Other, often overlooked, reasons for the slow transfer rate are the physical limitations. They could not be overcome at all. The minimal RTT (round time trip) between two servers is limited by the speed of light and usually does not exceed 2/3 of its value. This is a serious constraint particularly for cross continent transmissions. Let us suppose that 1 Kbyte of data is to be transmitted over a connection with a raw throughput of 2.48 Gbits. The average RTT between New York and Los Angeles is 85 ms. This translates into maximum data throughput of 4 Kbytes per second. The Figure 2 shows the relationship between the maximum throughput and the RTT.

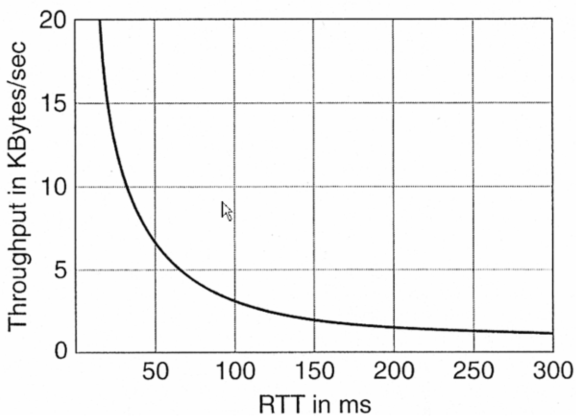


Figure 2. Theoretical TCP throughput for different RTTs for 1KB over a 2,48 Gps link [9]

That phenomenon is extremely worrisome because it means that even large improvements in the throughput have only marginal effect on transmission time of small objects.

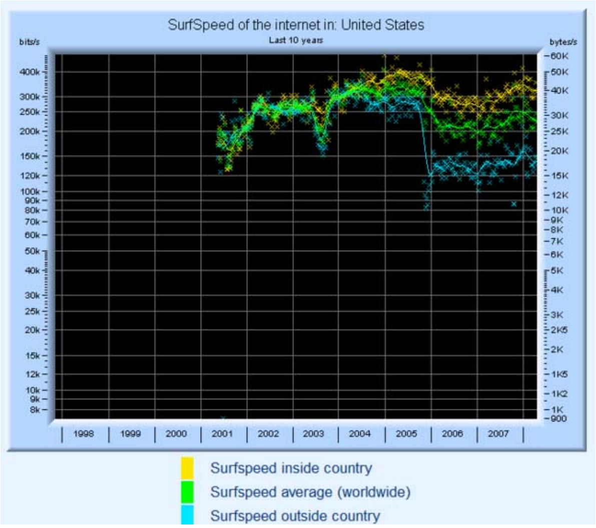


Figure 3. Surf speed in the USA [13]

Table 1. The pages used in the experiment

URL	Code	Origin	Size
ausdaily.net.au	AUSD	Australia	310
cheers.com.au	CHEERS	Australia	664
citynewsohio.com	OHIO	USA	158
faz.net	FAZ	Germany	1215
gazeta.pl	GAZETA	Poland	613
news.com.au	NEWS	Australia /CDN	1644
nypost.com	NYPOST	USA	1033
sueddeutsche.de	SUED	Germany	1198
suntoday.com.au	SUN	Australia	67
www.welt.de	WELT	Germany	989
www.wp.pl	WP	Poland	390

2. Browser Latency Test

The long time necessary to download Internet pages is the plaque of the WWW. Despite the constant increasing of the capacity of Internet network the actual transfer rate, the so called surf speed (described further in the 5th Section), does not show significant improvements. The Figure 3 shows the data for the USA.

Taking into account the increase in average page size we can hardly expect significant decrease of the download time. During the test the actual download duration of several front pages of websites were recorded. The websites were located on servers

that are spread all over the world. The download was performed periodically. The aim of the paper is to study the performance of various Internet connections and therefore the effects of caching was eliminated.

The Table 1 shows the basic properties of the tested pages. Their websites are of the same type (newspapers or news services) but are spread all over the world. The original server of one of them (NEWS) is located in Australia but is available to the users by AKAMI, the leading CDN company.

All of the pages are generated dynamically and their sizes change from one download to another. The column Size contains the mean value taken from several measurements of the total size – the value includes also all embedded objects.

The experiment was conducted on 3 test-beds:

- HM1 using a local residential IPS, the throughput of 1 Mbytes;
- HM2 using a nationwide ISP, the throughput of 1 Mbytes;
- UNI using a connection to Internet through a University network with a relatively high throughput of 100 Mbytes. It should be added however that the connection was shared by many users, e-learning servers and many others.

The nominal throughput for HM1 and HM2 test-beds was confirmed by net tool available on the Test Speed Webpage [11]. The slow dialup modems were excluded from the study as they are becoming less and less popular.

Table 2. Browser download durations

Page Code	**UNI**			**HM1**	**HM2**
	Mean	Min	Max	Mean	mean
AUSD	34	28	40	22	17
CHEERS	17	15	18	17	12
OHIO	9	9	10	9	8
FAZ	33	28	38	32	22
GAZETA	26	23	30	21	10
NEWS	42	36	48	25	31
NYPOST	67	58	77	26	25
SUED	15	13	17	16	16
SUN	2	2	4	2	1
WELT	18	17	20	16	15
WP	9	7	11	7	8

Surprisingly the tools for measuring the browser latency are not popular. The Charlotte package witch is both well developed and is widely available as it belongs to the freeware domain works only with outdated and no longer supported browsers [12]. For that reason a program was developed to collect data for the experiment. It uses the OLE to control the Microsoft IE.

The initial presumption was that the UNI test-bed having superior transfer rate should easily surpass the other competitors whereas the other two test-beds should have

roughly the same performance. The actual data contradicted the assumption to great extent. The download duration basic data is shown in the Table 2.

The data clearly indicate the usefulness of caching. Without it the 8 second limit was rarely achieved. The AKAMAI powered NEWS page is by far the largest but its download durations are still one of the longest.

In most cases the mean value for the UNI test-bed was surprisingly greater than the mean values for the other two ISPs. This is not a coincidence as the HM1 and HM2 means are outside the values calculated for confidence level α=0.05. The values are shown in the Min and Max columns. The mean download duration of 5 pages for HM1 and 9 pages for HM2 were significantly better than the values for the UNI. Not even in a single case the UNI test-bed was better than one of the other ISPs. The HM2 and HM1 test-beds were compared in the same manner. It turned out that on the same confidence level in 7 cases the durations of the HM2 were significantly shorter than those for the HM1 and only in one case the HM1 performed better. The HM2 test-bed is therefore the clear winner.

The mean is only one factor in the human assessment download duration. The other is the variation of the duration. People tend to forget about frequent short downloads and remember the few exceptionally long ones. The Table 3 shows Vs - the normalized standard deviation data calculated using the Eq. (1)

$$Eq.1 \qquad Vs = 100 \frac{std.dev.}{mean}$$

The variation is normalized in order to enable the comparison of pages with significantly different values of standard deviation.

Table 3. Normalized Standard Deviation of Download Durations

Page	**UNI**	**HM1**	HM2
AUSD	119.0	115.4	92.3
CHEERS	58.0	53.3	31.0
OHIO	27.0	42.2	26.7
FAZ	134.9	129.6	28.6
GAZETA	88.5	99.2	23.4
NEWS	51.0	38.7	20.6
NYPOST	77.1	40.7	15.9
SUED	56.9	46.4	31.3
SUN	140.8	139.4	59.3
WELT	43.7	55.2	25.7
WP	145.9	232.0	52.4
The Sum	943.4	992.0	407.0

Once more again the HM2 has achieved significantly better results than the other test-beds. In all three cases the top values are characteristic for the WP website. This is

one of the most popular news services in Poland and the high values reflect probably the facts that the server has problems with handling the surges of incoming download requests.

The download durations were also analyzed separately for three daytime periods: night, work and evening. The results are not published here but also in this case the HM2 test-bed has the most stable performance.

All data presented above contradict the initial presumption stated at the beginning of the section. The increasing of the transfer rate does not guarantee the shortening of download durations. In the following sections we will analyze other properties of the network connections in the individual test-beds trying to identify a property or properties that are correlated with the download durations.

Table 4. Ping Durations in milliseconds

Page Code	UNI			HM1	HM2
	mean	Min	Max	mean	mean
AUSD	184.7	183.4	186.0	232.8	206.8
CHEE	182.8	181.7	184.0	203.3	140.4
OHIO	121.8	121.3	122.4	143.0	64.2
FAZ	24.7	24.3	25.1	57.1	25.4
GAZE	2.9	5.4	40.3	30.2	29.6
NEWS	9.0	8.7	9.4	17.8	26.1
NYPO	9.2	8.7	9.7	18.8	52.4
SUED	30.2	30.0	30.5	47.6	185.3
SUN	159.7	158.7	160.7	175.0	62.5
WELT	44.0	43.5	44.5	51.4	156.7
WP	16.5	12.7	20.3	32.2	29.8

3. Low Level Measures

The low level measures analyzed in the section include: ping duration and trace route length and trace route duration. The ping and trace route data were simultaneously collected with the download durations. Analyzing the results one should take into account that both ping and trace route tools do not use the HTTP protocol.

3.1. Ping

Ping is a computer network tool used to test whether a particular host is reachable across an IP network. It works by sending ICMP “echo request” packets to the target host and listening for ICMP “echo response” replies. Although is does not measure the download time of a page, the Ping tool is widely used to estimate the RTT round-trip

time, expressed generally in milliseconds. As stated in the Section 2 the RTT time has a profound influence on the HTTP performance.

The Table 4 compares the ping durations of the UNI connection with the results of two HM connections. As usual the Table contains the mean PING duration together with the confidence interval.

This time the UNI test-bed is the clear winner, the HM2 coming in second place. Other conclusion could be drawn:

- the remarkable difference in the nominal bit rate does not match the differences in ping durations, which are not that different;
- the longer the distance to the original server the more uniform distribution of ping durations – compare the performance of the WP and AUSD servers;
- in all cases the best results has the AKAMI powered NEWS server.

Table 5. Similarity of Trace Route Hosts

Test bed	Page Code	UNI		
		a	B	C
UNIc	AUSD	0.72	0.34	1.00
	News	0.91	0.60	1.00
	WP	0.71	0.43	1.00
	Welt	0.72	0.22	1.00
	OHIO	0.78	0.59	1.00
HM1	AUSD	0.02	0.02	0.02
	News	0.02	0.02	0.02
	WP	0.12	0.14	0.09
	Welt	0.08	0.17	0.02
	OHIO	0.02	0.02	0.02
HM2	AUSD	0.26	0.12	0.29
	News	0.02	0.03	0.02
	WP	0.11	0.07	0.14
	Welt	0.15	0.00	0.19
	OHIO	0.03	0.03	0.03

3.2. Traceroute

Traceroute is a computer network tool used to determine the route taken by packets across an IP network on their path to destination server. The three timestamp values returned for each host along the path are the latency values in milliseconds for each packet in the batch.

Although the IP protocol does not guarantee that all the packets take the same route the test routes identified during the test indicate a significant amount of similarity. The Table 5 shows the similarity between sets of hosts obtained by 3 Traceroute traces

from the UNI test-bed (UNIa, UNIb and UNIc) and by traces in the other environments. The similarity level is measured by the Jaccard coefficient:

$$Eq.2 \qquad Sim(TrA, TrB) = \frac{|\ \overline{TrA} \ \cap \ \overline{TrB}\ |}{|\ \overline{TrA} \ \cup \ \overline{TrB}\ |}$$

where:

$\overline{TrX}$ is the set of host in the trace X

$|\ A\ |$ is the X cardinality of the set A.

Table 6. Number of Hosts in Trace Routes

Trace	Page Code				
	AUSD	NEWS	WP	WELT	OHIO
UNIa	21	8	10	19	22
Unib	22	16	30	63	63
Unic	61	49	30	19	22
Hm1a	23	6	14	18	23
Hm1b	23	6	14	17	24
Hm1c	23	6	14	18	23
Hm2a	15	18	7	12	14
Hm2b	20	15	6	12	14
Hm2c	15	4	6	17	14

Table 7. Duration in milliseconds of Trace Routes

Trace	Page Code				
	AUSD	NEWS	WP	WELT	OHIO
UNIa	1462	16	51	468	1199
Unib	1808	63	60	1344	4136
Unic	4943	176	152	467	1265
Hm1a	1527	38	334	487	1599
Hm1b	1427	40	256	460	1624
Hm1c	1399	45	276	514	1607
Hm2a	2785	638	233	749	1375
Hm2b	2024	663	125	545	1510
Hm2c	2266	72	97	1075	1514

The marked (shaded) figures show the similarity level of traces within a single test-bad (so called internal traces of the UNI test-bad) and they exhibit a fair amount of similarity. Although not shown in the table, the similarity of internal traces for other

environments was even greater. Truly astonishing is the extremely low similarity level between external traces (traces taken from different environments). Even for far away websites the level is in usually negligible. This is a clear indication that the raw throughput provided by the ISP could not be the decisive factor of the user perceived latency as the packets from different test-beds take strikingly different routs.

The next two tables show other characteristics of the traces: the number of hosts in each route (Table 6) and the cumulative latency (Table 7) reported by the tool.

Similar looking data were collected for the other pages. The low level of similarity between traces from different test-beds is easily to explain considering differences in number of their elements. A closer analysis of the data from the table reveals the following conclusions:

- the HM1 and HM2 environments show a remarkable amount of stability, both the number of hosts and the cumulative duration are similar;
- although the UNI connection is has often the lowest latency but is result are not as stable as for the other connections. Comparing with other test-beds many of its traces include more hosts and as a consequence the tool takes much more time to complete;
- as in the case of the ping utility the AKAMAI powered NEWS page can provide the shortest latency time.

Table 8. Average Surf speed in kilobits/sec

Test-bed	Mean	Min	Max	SD	Err
Connections within Poland					
UNI	1106	1013	1199	786	71
HM1	911	777	1045	926	102
HM2	238	198	277	245	103
Worldwide connections					
UNI	292	280	305	106	36
HM1	143	132	154	75	53
HM2	128	119	136	52	41

4. High Level Measures

The low level measures make use of different connection properties but are not directly connected with the user perceived latency or even with the HTTP protocol. They could be obtained by well known network utilities.

In this section a high level test – the so called surf speed is used. The test is provided by the mentioned above Numion server [13]. The test consists in downloading a number of objects with varying sizes from a set of popular servers. One can select countries to test. Unfortunately none of the used objects belonged to one of the tested websites.

In order to test the raw connection performance every precaution is taken to pass by all possible caches. It claimed that such a mode of operation is an accurate simulation of surfing and the measurement uses the computer exactly like a user would

and is influenced by exactly the same things affecting your surfing speed, such as programs running in the background, hardware (video, hard disk), and the efficiency of your browser. The actual download duration is sent to the Numion server. The server makes available a wide range the statistics enabling the users to identify trends in both their local environment and on the country or even world level. This is a popular service, the number of separate measurements taken each week world wide exceeds well 100 thousands and the total number is over 64 millions (April, 2008). The data is summarized in the following tables. The Table 8 contains 24 hour data whereas the Table 9 shows data separately for night, work, and evening.

Table 9. Surf speed for day periods

Test Bed	Day Period	Mean	Min	Max	Std Dev	Err
Connections within Poland						
UNI	n	1391	1176	1606	710	51
UNI	w	1147	1022	1272	814	71
UNI	e	841	679	1002	690	82
HM1	n	869	428	1311	843	97
HM1	w	888	738	1039	916	103
HM1	e	1046	665	1428	1031	99
HM2	n	229	179	279	201	88
HM2	w	229	166	291	206	90
HM2	e	259	164	354	326	126
Worldwide connections						
UNI	n	388	356	419	104	27
UNI	w	303	292	314	73	24
UNI	e	211	184	238	114	54
HM1	n	130	106	155	47	36
HM1	w	135	123	147	72	53
HM1	e	193	160	225	87	45
HM2	n	126	117	136	39	31
HM2	w	128	109	146	61	48
HM2	e	130	112	147	60	46

The Tables 8 and 9 contain the Mean, Min, Max columns known from the previous tables. The additional SD and Err columns represent the standard deviation and normalized error respectively. The results are surprising. One could reasonably expect that a high level measure such as the speed test should provide results similar to the results of the browser latency test. It is definitely not the case. The best in the browser test, the HM2 test-bed is the slowest one in the Numion test, especially when the

connections to servers in Poland are concerned. Please note the scale of the difference. It is really remarkable. The mean surf speed is more than 4 times less than the speed for the UNI test-bed. The difference between the results of the UNI and HM1 environments, although statistically significant, is not as great as could be inferred from the difference in the raw network throughput of both environments. Other results are more predictable. In all cases the worldwide servers are slower than the local servers. The leader is the UNI environment connection with highest throughput and lowest variation.

Table 10. Input data for the Spearman test

Page	Ping	TL	TD	#	1Kb
			UNI		
AUSD	185	30	2950	100	0.11
OHIO	122	26	1555	15	0.06
FAZ	25	17	207	147	0.03
GAZ	23	13	64	163	0.04
NEWS	9	19	76	174	0.03
SUED	30	18	323	101	0.01
WELT	44	23	548	113	0.02
WP	17	12	65	117	0.02
			HM1		
AUSD	233	23	1500	100	0.07
OHIO	143	24	1641	15	0.06
FAZ	57	16	147	147	0.03
GAZ	30	7	57	163	0.04
NEWS	18	6	38	174	0.02
SUED	48	13	331	101	0.01
WELT	51	17	474	113	0.02
WP	32	14	290	117	0.02
			HM2		
AUSD	207	17	2970	100	0.05
OHIO	64	14	1537	15	0.05
FAZ	25	12	1401	147	0.02
GAZ	30	6	278	163	0.02
NEWS	26	13	628	174	0.02
SUED	185	8	647	101	0.01
WELT	157	14	894	113	0.02
WP	30	6	174	117	0.02

The surf speed of the UNI test-bed is not only the highest but also the most stable one. The HM1 test-bad shows better performance than the HM2 test-bed but it is statically significant only during the evening period. For the rest of the day the difference is not statistically significant.

5. Relevance of Tests

During the experiment a number of measures were used. In our opinion the browser test simulates best the surfing experience of a regular user. The measure is difficult to apply. It requires a careful selection of tested Web pages and could not be obtained by the standard network tools. The aim of the section is to identify which measure provides results most similar to the browser test. The similarity is calculated using the Spearman coefficient. The coefficient compares the various rankings of tested pages, separately for each environment. The input data is shown in the Table 10. The pages were ranked according to the mean values of: ping duration, trace route length (column TL), trace route duration (TD), number of objects of a page (#), and the transfer rate reported by the browsing speed (1 Kbyte) – it specifies the time is sec for transmitting 1 Kbyte of data. The transfer rate replaced the download latency because this time we compare different pages within the same test-bed so one has to make up for the differences in page sizes. The average number of objects in consisting of page are included for the reference.

In the next step the rankings were compared using the Spearman coefficient. The raw scores are converted to ranks, and the differences d_i between the ranks of each observation are calculated. The coefficient is calculated according to the following formula:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where:

- d_i: the difference between a rank in corresponding orderings.
- n: number of pairs of values.

The value of the coefficient is always in the range from -1 to 1. The greater is its absolute value the more significant is the correlation. Negative values indicate that the increase of value of one property is correlated with the decrease of value of the corresponding second property.

The base ranking was defined by the 1 Kbyte columns. The values of the Spearman coefficient for all other columns are given in the Table 11.

Table 11. Values of the Spearman coefficient for low level measures

Test-bed	Ping	TL	TD	#
UNI	0.381	0.405	0.286	-0.214
HM1	0.619	0.619	0.429	-0.333
HM2	0.071	0.405	0.381	-0.310

For all three test-beds the highest correlation level was found between the browser speed and the trace route length. The measure performed even significantly better than

the trace route duration which is a little surprising. The values for the number of objects are negative which is not surprising. The Ping values are useful for the HM1 test-bad but are totally useless for the HM2 test-bed.

6. Conclusions and Further Area of Study

The aim of the reported study was to discover in what way the Internet connection influences the download latency. The popular assumption is that a raw connection throughput is the determining factor in browser latency. The notion is by obvious reasons promoted by the ISPs and it is also keeping with the common sense. The experiment described in the paper indicates that such an assumption is not valid. The influence of the trace route path is far greater. As far as we know that phenomenon was not reported previously. The IP protocol is the reason of the unpredictability of the routes taken by packets and this has probably discouraged research on the area. The obtained results suggest however that there is a fair amount of stability in the trace routes. This gives hope that further study on the area could produce meaningful results.

The scope of the performed experiment does not allow us to identify why the speed-surf test and the browser latency test produced different results. This may lay in the fact that the test used different WWW servers. It has to be confirmed by further study in which browser latency test downloads pages from the servers used in the surf-speed test.

Another interesting subject of study is the usefulness of the CDN services. The only one server that used the AKAMAI CDN service (page NEWS) had excellent values of low level test. The NEWS page was at the same time one of the most sluggish pages in the browser latency test. This is in part due to the remarkably great size of the page but the result is still somehow disappointing.

The collected data are adequate to evaluate the significance of different factors but are not sufficient enough to propose and verify a model of an Internet connection. Such a model must encompass several factors. Its aim would be to rank the Internet connections using a small set of well defined measures. A simple model was proposed in [14]. The model was, however, not verified in any way and it contains hard to measure factors such as original server processing time. It looks however as a good starting point for a further study. The precondition is the availability of a far larger pool of input data. Such a data set is being currently collected.

References

[1] N. Bhatti, A. Bouch, A. Kuchinsky, Integrating User-Perceived Quality into Web Server Design, *Computer Networks* 33 (1999).

[2] S. A. Card, T. P. Moran, A. Newell: *The psychology of Human-computer Interaction*, Lowrence Erlbaum Associates, NJ, 1983.

[3] J. Ramsay, J. Barbesi, I. Preece, Psychological Investigation of Long Retrieval Times and the World Wide Web, *Interacting with Computers*, 10 (1998).

[4] Zona Research, *The economic impacts of unacceptable Web Site download speeds*, White paper, http://www. zonaresearch.com/deliverables/while-apers/wp17/index.html.

[5] C. Cairano-Gilfedder, R. G. Clegg, A decade of Internet research - advances in models and practices, *BT Technology Journal* archive 23 (2005), 115-128.

[6] A. Siemiński, Changeability of Web objects - browser perspective, *5th International Conference on Intelligent Systems Design and Applications*. Proceedings, Wrocław, September 8-10, 2005, IEEE Computer Society [Press], cop. 2005. 476-481.

[7] A. Siemiński, Browser latency impact factors. *Lecture Notes in Computer Science, Lecture Notes in Artificial Intelligence* 4693 (2007), 263-270, 11th International Conference, KES 2007.

[8] P. Destounis, J. Garofalakis, P. Kappos, Measuring the Mean Web Page Size and Its Compression To Limit Latency and Improve Download Time, *J. Internet Research* 11 (2001), 10-17.

[9] M. Rabinowich, O. Spatschech, *Web Caching and Replication*, Addison Wesley, USA, 2002.

[10] Z. Wang, P. Cao, *Persistent Connection Behavior of Popular Browsers*, http://pages.cs.wisc.edu/~cao/papers/ persistent-connection.html.

[11] *http://www.speedtest.net/index.php*.

[12] L.R. Lamplugh, D. Porter, Charlotte, An Automated Tool for Measuring Internet Response Time, *Novell research*, 1998.

[13] http://numion.com.

[14] A. Savoia, Webpage response time. *http://www.stqemagazine.com/*, July/August 2001, 48-53.

This page intentionally left blank

Subject Index

New Trends in Multimedia and Network Information Systems
A. Zgrzywa et al. (Eds.)
IOS Press, 2008

Author Index

This page intentionally left blank

This page intentionally left blank

This page intentionally left blank